Characterizing the Novel Roles of Small RNA-directed DNA Methylation in the Epigenetic Regulation of *Athila* Family Retrotransposons

Honors Research Thesis

Presented in Partial Fulfillment of the requirements for graduation with honors research distinction in Molecular Genetics in the undergraduate colleges of The Ohio State University

By

Erica Thomas

The Ohio State University

April 2014

Project Advisor: Professor R. Keith Slotkin, Department of Molecular Genetics

Table of Contents:

Background

Transposable elements (TEs) are sequences of DNA that possess the ability to mobilize from one location of a genome to another. When active, TEs duplicate themselves to high levels. Because of their ability to duplicate, TEs have accumulated in the genome of every eukaryote and some prokaryotes. For example, 48% of the human genome is comprised of TEs [11]. TE activity is extremely mutagenic, as mobilization results in chromosome breaks and insertions into essential genes, severely disrupting the sequence of the genes' codons. TEs are known to generate many mutations in the human genome [1]. Mutations produced by TEs are frequently the cause behind diseases such as hemophilia A and B, severe combined immunodeficiency, and Duchenne muscular dystrophy [7]. Various TE-induced mutations can also lead to a progression of cancer [7]. Because every eukaryotic organism, including humans, carries TEs in their genome, TEs have been called "the mutagen from within".

There exist two classes of TEs within the genome. Class I elements, known as retrotransposable elements, transpose via a copy-and-paste mechanism of replication, which allows for their high levels of accumulation throughout the genome, potentially causing great damage to the host cell. This process begins when the retrotransposon synthesizes an RNA copy of itself. This RNA intermediate travels to a different part of the host genome, where it is then copied into DNA and reinserted into the genome. Class I elements can be further divided into Long Terminal Repeat (LTR) and non-LTR elements. LTR retrotransposable elements have a structure similar to LTR retroviruses, and it is likely that these retrotransposons were indeed once retroviruses which lost the ability to exit the cell at some point in their evolutionary history. Class II elements, or DNA elements, transpose without the use of an RNA intermediate via a cut-and-paste mechanism, potentially causing detrimental double-stranded breaks upon excision and disruptions of protein-coding sequences.

The focus of this research proposal is on a family of retrotransposons called *Athila*, found in the genome of the reference plant and powerful model organism, *Arabidopsis thaliana*. Arabidopsis is a small plant that many researchers use to increase the speed and efficiency of their research, because the molecular and genetic process related to silencing TEs are highly conserved in plants and mammals. The multi-copy *Athila* retrotransposon family composes more than three percent of the entire Arabidopsis genome, making it the largest family of TEs in Arabidopsis [15]. *Athila* is classified as a long terminal repeat (LTR) retrotransposon, and as such it is related to LTR retroviruses, such as Human Immunodeficiency Virus (HIV), but unlike retroviruses, *Athila* and other retrotransposons are unable to leave the cell to infect another cell or organism. The structure of a specific type of *Athila*, *Athila6* is shown in Figure 1.

Because *Athila* is a retrotransposable element, its replication cycle depends on the synthesis of an RNA intermediate, which is performed by the endogenous Polymerase II (Pol II) of the Arabidopsis host cell. Pol II begins transcription of the RNA intermediate at a promoter site in the 5'LTR region. This transcript eventually leads to the production of the TEs' own proteins which are necessary for its replication. Two proteins produced by *Athila*, reverse transcriptase and integrase, are encoded by its pol gene. The reverse transcriptase generates a DNA copy of the RNA intermediate, which is then inserted into the genome by the activity of integrase. Reverse transcription takes place within the other protein produced by *Athila*, known as the gag capsid particle, which is encoded by its *gag* gene. The replication cycle of retrotransposons differs from retroviruses in that retrotransposons do not encode a functional envelope protein, which is a necessary component for the cellular exit strategy of retroviruses.

In order to maintain a stable and mutation-free genome in the presence of the mobile and mutagenic TEs, eukaryotic cells possess a mechanism called epigenetic regulation to repress and silence the TEs. Epigenetic regulation is interesting because once triggered for silencing, the TE

sequence can be passed from generation to generation in this silenced state without the need to re-trigger the silencing after each cell division or generation. The term epigenetic regulation encompasses all of the ways in which a change in gene expression can be inherited without an alteration in the underlying DNA sequence.

Two pathways in particular work to ensure that TEs are epigenetically silenced and unable to cause significant harm to the genome. The first pathway, called ***RNA-directed DNA Methylation (RdDM)***, works to establish and maintain the cytosine DNA methylation of TEs (Figure 2). RdDM works as a transcriptional gene-silencing pathway, because the resulting DNA methylation established by RdDM prevents TEs from transcribing their mRNA and producing the proteins needed for transposition. The pathway begins with a plant-specific polymerase known as Pol IV, which produces a non-coding nuclear RNA transcript from TE DNA. This transcript is made double-stranded by RNA-dependent RNA polymerase 2 (RDR2). A hallmark of the activity of this pathway in the generation of 24 nucleotide siRNAs by the Dicer-Like 3 protein (DCL3). The 24 nucleotide siRNAs are then incorporated into AGO 4, 6, or 9, which uses the siRNA to bind to a scaffold transcript produced by another plant-specific polymerase, Pol V. AGO attracts cellular machinery which methylate and modify the histones and the DNA [10].

If the TE is transcribed into mRNA by Polymerase 2 (Pol II), the second pathway, called RNA interference (RNAi), degrades the RNA of expressed TEs to prevent TE protein production and mobilization (Figure 3). When expressed, the TE transcript is made double-stranded by RNA-dependent RNA polymerase 6 (RDR6). The dsRNA is cut into 21 or 22 nucleotide siRNAs by the Dicer-Like 4 or 2 proteins respectively (DCL4 & DCL2). The siRNAs are then incorporated into Argonaute 1 (AGO1), which uses the siRNA to bind to and target TE RNA for degradation or tranlational repression [14]. Since this pathway is functioning after the TE has already transcribed itself,

it is referred to as a post-transcriptional gene-silencing pathway. Without both the RdDM and RNAi pathways working to silence TEs, the host organism is left vulnerable to their mutagenic effects.

In order to observe TEs in their active state, our lab uses a mutant which is deficient in the activity of the protein Decrease in DNA Methlyation 1 (DDM1). DDM1 encodes a SWI2/SNF2-like protein. SWI2/SNF2-like proteins are involved in a variety of functions including transcriptional control, DNA repair, chromosome folding, and chromosome regulation. DDM1 acts as a chromatin remoldeler by hydrolyzing ATP to rearrange nucleosomes into more compact formations, which creates transcriptionally silent heterochromatin. DDM1 silences TEs by modifying the nucleosomes encompassing the TE DNA [9]. Without DDM1, the cell cannot condense chromatin into heterochromatin, and TE silencing cannot be maintained. In *Arabidopsis,* there is a 70% decrease in DNA methylation in *ddm1* mutants, and TEs become transcriptionally active [6]. Throughout the course of this report, several lines of *ddm1* mutants are used. One line, known as SL001, are plants which are second generation *ddm1* mutants. Another commonly used line is SL300, which consists of sixth generation *ddm1* mutants. These different lines are utilized in order to observe the effects that TE activity has over multiple generations.

There are several classes of DNA methyltranferases which work to place methyl groups on cytosine residues of TE sequences. One class is known as *de novo* methyltransferases, which create new methylation on DNA. *De novo* methylation can be recognized because it occurs in all sequence contexts. This includes CG, CHH, and CHG methylation, where H represents any base pair except guanine. The protein known as Domains Rearranged Methylase 2 (DRM2) is the methyltransferase which functions to create *de novo* methylation in Arabidopsis, and is targeted by RdDM [16]. The second class of methyltransferases recognize existing methylation on the parental strand of DNA, and establishes new methylation on the daughter strands, thereby maintaining methylation. This

maintanence methylation can be recognized because it occurs mostly in the symmetrical context on CG and CHG sites. The maintenance methyltransferases in Arabidopsis include DNA (cytosine-5)-methyltransferase 1 (MET1), which is responsible for CG methyaltion, chromomethylase 3 (CMT3), which is responsible for CHG methylation, and chromomethylase 2 (CMT2), which is responsible for CHH methylation [8]. Both MET1 and CMT3 function at high levels compared to CMT2, which is why the levels of CHH methylation are much lower in a maintenance context than both CHG and CG methylation.

Primary Focus of Project

The following thesis project has two main objectives:

1) **Refine the molecular characterization of the retrotransposable element *Athila6*.**

In order to begin to understand how *Athila* is epigenetically regulated by Arabidopsis cells, more information about the molecular characteristics of *Athila* must be established. Certain information, such as an understanding of the start and stop sites of transcription, is needed to perform further experiments of interest. Although many of the molecular features of *Athila* have been well established, several specific molecular characteristics of *Athila* remained unknown. A particular member of the *Athila* family, known as *Athila6*, is used throughout the course of these experiments because it produces the most amount of small RNAs, and therefore is presumably the most active of the *Athila* family members. In this thesis, I identify the following in *Athila6*:

a) An intron in the 3'region

b) The major and minor 5' transcriptional start sites of its transcript

c) The start site of a transcript which begins in the intergenic region between *gag/pol* and *env*

d) The 3' end site of the transcripts

2) **Explore in detail the potential relationship between the roles of RNA-directed DNA Methylation and endogenous RNA Interference in the transcriptional silencing of TEs.**

Previous research that was performed by our group suggested that the two pathways of RdDM and RNAi, which were thought to function separately, may both play a necessary part in the complete transcriptional silencing of TEs. This was realized when our lab mutated certain components of RNAi in a *ddm1* mutant background (TEs transcriptionally active) and observed a loss of *de novo* methlyation at the *Athila6* promoter, as well as an increase in *Athila6* expression [12]. Because of the dependence of the DNA methylation on RDR6, we have termed this pathway RDR6-RdDM. We have found that together, the previously known Pol4-RdDM pathway and the newly discovered RDR6-RdDM pathway have an additive effect in the corrective reestablishment of TEs. Furthering the investigation of this new pathway, I explore the following:

a) **The additional components of RDR6-RdDM**. I first worked to identify the specific Argonaute protein involved in this process. I analyzed the expression levels of each of the ten AGOs in Arabidopsis when TEs are active in a *ddm1* background to see if there is any upregulation of AGO steady-state mRNA levels in the presence of active TEs. I then analyzed expression levels of *ddm1/ago* double mutants for each of the ten AGOs in order to see if there is any significant increase in *Athila* expression with the loss of a particular AGO. In addition, I also analyzed levels of *de novo* methylation at the *Athila* transcriptional start site (TSS) in these mutant backgrounds. After that, I performed a small RNA Northern in several of the *ddm1/ago* double mutants to look for any changes of the small RNA profiles in these backgrounds. I also analyzed the *Athila* expression and DNA methylation levels

in *ddm1* double mutants for other non-Argonaute genes that we suspect are involved in the RDR6-RdDM pathway. These proteins include the following: DRM2, NERD, and SDE3. DRM2 is the only *de novo* methyltransferase known to function in Arabidopsis [16]. NERD is a protein needed for RDR2-independent DNA methylation [13]. SDE3 encodes an RNA helicase which is required for post-transcriptional regulation of TEs [2]. I then performed a TE display for a *ddm1/rdr6/polIV* triple mutant to see if enough repression of TE activity is lost that the TE is allowed to fully transpose to a new location in the host genome.

b) **The role of RDR6-RdDM in the resilencing of transcriptionally active TEs**. It has been shown in previous literature that when a wild-type plant inherits active TEs which are generating high levels of small RNA from one parent and silenced TEs from the other parent, the resulting offspring will have silenced normal TEs [4]. In order to further investigate the possibility that RDR6-RdDM is involved in this resilencing process, I generated a line of plants which are heterozygous for *ddm1* and homozygous mutant for particular RDR6-RdDM components. These components include RDR6 and DCL2. I also included components from the Pol4-RdDM pathway to compare the effects of the loss of this pathway and the loss of RDR6-RdDM. The Pol4-RdDM components analyzed in these experiments are RDR2, DCL3, Pol4 and Pol5. I then analyzed *Athila* expression levels using qRT-PCR as well as DNA methylation levels at the promoter region of *Athila* using bisulfite sequencing to see if there is a loss of *de novo* methylation, therefore resulting in a higher accumulation of TE transcripts. If *Athila* expression levels increase due to a decrease in methlyation levels at the *Athila* promoter when components of RDR6-RdDM are not functional, then this suggests that RDR6-RdDM is important in the transcriptional resilencing of TEs.

c) **The potential of RDR6-RdDM to resilence TEs beyond *Athila***. In order to explore this idea, I performed the resilencing investigation previously explained on the DNA TE known as

ENSPM5.

d) **Identify the time point in the development of Arabidopsis in which the resilencing of TEs occurs**. In order to pinpoint the exact point in developmental time in which RDR6-RdDM is working to resilence TEs, I analyzed the *Athila* expression levels in the following tissue types representing different points in the development of Arabidposis: seed, seedling, juvenile leaf and inflorescence tissue. If the exact point in developmental time in which the resilencing of TEs occurs could be identified, then this particular tissue type would be utilized in future investigations of the activity of RDR6-RdDM.

Frequently Utilized Techniques

1) **5'RACE System for Rapid Amplification of cDNA Ends**. 5'RACE is used to identify the exact start site of mRNA transcripts. It begins with the isolation of RNA from each sample of interest, followed by the selective degradation of non-mRNA. Then, the 5'Cap structure which exists on all mRNAs is removed and an oligo of known sequence is ligated onto the decapped mRNA. The oligo capped 5'RNA is then converted into cDNA using an oligo-dt primer and the enzyme *reverse transcriptase*. The products are then amplified using two rounds of nested PCR which utilize two different sets of primers. Each set of primers includes one primer which is specific to the oligo adapter sequence, and another primer specific to the 5'region of the transcript of interest. The results of these PCRs are analyzed on an agarose gel. The bands of interest are gel extracted, cloned, and then sequenced in order to identify the exact base pair of the transcriptional start site.

2) **3'RACE System for Rapid Amplification of cDNA Ends.** 3'RACE begins with the conversion of mRNA to cDNA using the enzyme reverse transcriptase, a primer which is specific to the 3'region of

interest, and an adapter primer which is specific to the poly-A tail that exists on all mRNA transcripts. The products are then amplified using two rounds of nested PCR using two different sets of primers. Each set of primers includes one primer which is specific to the 3'region of the target transcript, and another which is specific to the poly-A tail of the transcript. The results of these PCRs are analyzed on an agarose gel. The bands of interest are then gel extracted, cloned, and then sequenced in order to identify the exact base pair of the poly-A addition site.

3) **Bisulfite Sequencing**. In order to identify DNA methylation levels, bisulfite sequencing must be used. The first step in this set of experiments is to isolate DNA from each tissue of interest, and then RNase these samples in order to degrade any RNA that may be left after the DNA extraction. This RNase-treated DNA is used in a bisulfite conversion reaction which converts unmethylated cytosine residues to uracil while leaving 5-methylcytosine residues unaffected. The bisulfite treated DNA is then amplified using primers specific to the target of interest. As a negative control, the unmethylated start site of a control gene is also amplified. The PCR products are then purified and sequenced, and this sequence is then compared to a reference consensus sequence using a software program known as Kismeth [5]. This comparison performed by the program calculates the amount of converted and unconverted cytosines to the consensus sequence, and gives an output of the various types of methylation found on the sequence, including CG, CHG, and CHH.

4) **Quantitative Reverse-Transcriptase Polymerase Chain Reaction (qRT-PCR)**. qRT-PCR allows for the quantification of the amount of cDNA in a sample of interest. In order to perform qRT-PCR, RNA must first be extracted from the samples of interest, and the samples must be DNased to remove any DNA which could cause contamination issues in future steps. The remaining RNA is then made into cDNA utilizing the enzyme reverse transcriptase. This cDNA is then added to primers specific to the target site of interest, SYBR Green fluorescent dye, and Taq polymerase. The cDNA is then

denatured, and the primers are annealed to the now single-stranded cDNA. As Taq adds nucleotides and extends the template, the SYBR Green fluorescent molecules are incorporated into the newly-formed double stranded DNA. SYBR Green only fluoresces when it is situated in double-stranded DNA. This allows for the fluorescence to represent the relative amount of DNA, which is measured by the qPCR machine.

5) **Small RNA Northern.** Small RNA Northerns are used in order to identify relative amounts of different size classes of small RNAs (sRNAs) in a given background. To begin this technique, small RNAs are extracted and purified from each of the samples of interest. These sRNAs are then run out on a polyacrylamide gel in order to separate the different size classes of sRNAs. This gel is then transferred and crosslinked to a neutral membrane, and then allowed to dry. The blot is then probed with radioactivity, allowing for sRNA presence and levels to be detected.

6) **Transposable Element Display (TED)-** Amplified fragment length polymorphism (AFLP) Transposable Element Display is used to detect new TE insertions. TED begins with the random digestion of the genome, followed by the ligation of these genomic fragments onto an adapter of known sequence. Then, these fragments are amplified twice, each time using one primer specific to the adapter and another primer specific to the *Athila2* or *6* sequence. Since the genome was randomly digested into different sized fragments, each *Athila* element will amplify as a different sized fragment. The secondary PCR products are radioactively labeled, and run out on a polyacrylamide gel. New insertions are detected by restriction site polymorphisms generated upon new TE insertion. These polymorphisms are depicted as new radioactive bands on the polyacrylamide gel.

Molecular Characterization of *Athila6*

Materials and Methods

In order to begin to uncover the details of *Athila*, the goal of my first project was to identify the precise location of an intron that was suspected to exist in *Athila's* sequence. The existence of an intron was suspected because of the difference in size between *Athila* DNA and cDNA, which was first observed by a previous undergraduate in the lab, Jennifer Bosse. To do this, I first extracted DNA from wild type *Arabidposis thaliana* in the Columbia ecotype (Col), *ddm1* mutants, and *ddm1/rdr6* mutants. I then performed polymerase chain reaction (PCR) using extracted inflorescence DNA and primers which flanked the suspected intron region. I also performed RT-PCR on cDNA made from inflorescence RNA using the same primers. The size of the cDNA in mutants expressing *Athila* was then compared to the size of the wild-type DNA by running both the DNA and cDNA products on an agarose gel. In order to identify the exact location of the intron, the amplified cDNA and DNA products were the cloned, sequenced, and then compared to the *Athila* consensus sequence using Basic Local Alignment Search Tool (BLAST).

The next set of experiments to characterize *Athila* involved the identification of the start site of its transcripts. The precise sequence at the 5' end of a sequence was determined by 5' RACE. RNA was isolated from *ddm1/rdr6*, *ddm1/dcl4*, and the *ddm1* mutants SL001 and SL300. The *ddm1/rdr6* and *ddm1/dcl4* mutants were included to explore the possibility that *Athila* could produce a transcript beginning at an alternative start site when post-transcriptional degradation of transcripts is compromised. Columbia was used as a negative control to ensure that there was no contamination of *Athila* transcripts. Multiple PCR bands were gel extracted, cloned, and sequenced.

I also set out to explore the possibility that *Athila* was producing a transcript with a start site around its ancient *env* region. Once again, I used 5'RACE to identify this potential start site, but this

time I utilized primers specific to the intergenic region 5' of the *env* region. The backgrounds utilized in this set of experiments include *dcl2/4*, *dcl2*, *dcl4*, Col, and *ddm1*. The *dcl2/4*, *dcl2*, and *dcl4* mutants were included in this analysis because our group generated previous data that hinted towards the expression of *Athila* transcripts that were shorter than the full length transcripts beginning in the 5'region in these particular mutant backgrounds. Multiple bands labeled were gel extracted, cloned, and sequenced.

Finally, I identified the 3' end site of *Athila* transcription. This was done by utilizing 3'RACE PCR. The samples included in this analysis are *ddm1* F2 (SL001), *ddm1* F6 (SL300), *ddm1/dcl4*, and *ddm1/rdr6.* The *ddm1/rdr6* and *ddm1/dcl4* mutants were included to explore the possibility that Athila could produce a transcript ending at an alternative end site when post-transcriptional degradation of transcripts is lost. Two bands which were very close in size were observed, and these bands were gel extracted, cloned and sequenced to determine the exact end site of *Athila* transcription.

Data and Analysis

The results of the investigation of the existence of an intron in the 3'region of the *Athila* sequence are shown in Figure 4. This figure indicates that there is indeed a size difference in the DNA and cDNA of *Athila6* in its 3'region, supporting the existence of an intron. The figure also indicates that *Athila6* is producing both spliced and unspliced transcripts, because a band exists in cDNA lanes that appear to be the same size as the *Athila6* DNA. When the cDNA of the transcripts were sequenced and compared to the *Athila6* DNA sequence, they were either not missing any nucleotides (full length transcript indicated by the arrow), or missing approximately 1,810 base pairs close to the *env* and 3'LTR regions of the element (Figure 5).

In the investigation of the 5'RACE TSS of *Athila6* shown in Figure 6, two bands were observed

in *ddm1* F6 (bands A and B), while one band was observed in *ddm1* F2 (band C). When these bands were gel extracted, sequenced, and then BLASTed against the *Athila* consensus sequence, the exact start site of the transcripts could be determined. Bands A and C were shown to have a start site around base pair 725 of the *Athila* 5'LTR. This represents the major start site of *Athila* transcription in the genome. Band B exhibited a start site around base pair 917 of the *Athila* 5'LTR. This second start site is a minor start site because only one specific *Athila* element in the Arabidopsis genome (At1g40095) shares sequence homology with the sequences cloned from this band (Figure 7).

The *env* region of *Athila* was also suspected to produce a transcript. This is because the ancient transcript would have encoded the envelope protein needed for *Athila* to exit the cell when it was still functioning as a retrovirus. Even though this transcript is no longer serving to produce a protein, it is still likely that it is being produced at a different promoter than the one located in the 5'LTR. The genotypes utilized in this experiment include Col, *dcl4*, *dcl2*, *dcl2/dcl4*, and *ddm1*. These genotypes were selected because it was previously shown by our group that 21/22 nt siRNAs of this particular region of *Athila* were being produced in Col, *dcl2,* and *dcl4,* hinting towards the activity of *Athila* in these backgrounds. The *env* start site data showed similarly sized bands in Col, *dcl4*, and *ddm1*. A larger band was observed in the *dcl2* background (Figure 8). The *Athila* consensus sequence used to identify the exact location of the start sites begins with the last 492 base pairs of the *gag/pol* region, and ends at the start of the *env* region, approximately 1825 base pairs later. When the bands were gel extracted, sequenced, and then BLASTed against the *Athila* consensus sequence, the larger band in *dcl2* was shown to have a start site approximately 350 base pairs into the end of the *gag/pol* region, and the smaller bands in Col, *dcl4*, and *ddm1* matched to the consensus sequence at around base pair 1040 in the intergenic region (Figure 9).

The poly-A site data of *Athila6* showed two very similarly sized bands, indicating that the two

different end sites are very close to one another in the consensus sequence (Figure 10). The genotypes utilized in this experiment include Col, SL001 (*ddm1* F2), SL300 (*ddm1* F6), *ddm1/dcl4*, and *ddm1/rdr6*. The *ddm1/rdr6* and *ddm1/dcl4* genotypes were selected because there is more *Athila* expression in these backgrounds due to the loss of endogenous RNAi. When these bands were gel extracted, cloned, and sequenced, the end sites matched to the consensus sequence at around an average of 613 base pairs into the 3'LTR region (Figure 11). A complete figure of *Athila* including its intron, start sites, and end site is depicted in Figure 12.

Conclusion

The data from these sets of experiments helped to further the basic characterization of *Athila*. Through these experiments I have shown that there are two different transcripts produced by *Athila*, one of which is spliced and the other unspliced. Future work can be done to understand the purpose of the production of a full-length and spliced transcript, as well as any differential regulation that may occur between the two transcripts. Also, I have confirmed the approximate location of the intron, as well as its length, which is an average of 1,810 base pairs in length.

In addition to the identification of the intron in *Athila*, I have also identified the specific regions in which the start site of its transcripts begins. There exists one major TSS in the 5'LTR that starts at approximately base pair 725 of the consensus sequence, while another minor TSS begins at base pair 917 of the consensus. The minor TSS is only observed in one *Athila6* element known as At1g40095. It is likely that this particular element has experienced a mutation at some point throughout evolutionary time that altered its start site. The major 5'LTR start site serves as the target region for all the bisulfite sequencing performed in the remainder of this thesis. *Athila* also possesses several other transcriptional start sites around its *env* region. Starting with the last 492 base pairs of the *gag/pol* region and

continuing to the start of the *env* region (base pair 1825), *dcl4*, *ddm1*, and Col all exhibit start sites around base pair 1040 in the intergenic region. 5'RACE is a very sensitive assay which amplifies the cDNA many times throughout several rounds of PCR. It is likely that there are very low levels of *Athila* expression in *dcl4* and Col when compared to *ddm1*, but these low transcript levels are amplified to detectable levels. However, even if these *Athila* transcript levels are low in both *dcl4* and Col, a future goal would be to understand why these transcripts are being produced in these backgrounds where they were previously thought to be silent.

Additional future directions of the molecular characterization of *Athila* include investigating the possibility that the 5'LTR transcript and the *env* transcript have different 3'end sites. Also, it is still unknown if the *env* transcript possesses the 3'intron found in the 5'LTR transcript, or if it remains unspliced. This could be investigated by performing RT-PCR of this particular region of the transcript and comparing it to the DNA sequence of that same region to see if there is any size difference between the sequences.

The Additional Components of RDR6-RdDM

Before our group began to investigate the RDR6-RdDM pathway, it was previously thought that the components of RNAi were solely involved in the post-transcriptional regulation of TEs. However, it was shown that the methylation levels of the *Athila* promoter were significantly reduced in a *ddm1/rdr6* background. Also, *Athila* transcript levels increased in a *ddm1/rdr6* background when compared to *ddm1* (Figure 13). This was an unexpected result, and led to the investigation of the role of RDR6 and other RNAi components in the transcriptional regulation of *Athila* through a pathway which our group has named RDR6-RdDM. This is an important pathway to understand because it is another way in which Arabidopsis has evolved to silence mutagenic TEs.

Materials and Methods

In order to determine the proteins that are acting in the newly identified RDR6-RdDM pathway involved in the *de novo* methylation of TEs, I generated multiple double mutant lines, which consisted of *ddm1* mutants and the mutated protein that was suspected to be working in the RDR6-RdDM pathway. The first set of proteins that I investigated was the family of Argonaute proteins, which incorporate small RNAs and use their sequence homology to target various RNA and recruit regulatory machinery to these regions. The mutant lines included nine Argonautes in the Arabidposis genome, excluding *ddm1/ago5* because of complications that were encountered when generating the line. I included a strong and a weak allele for AGO1 and AGO9, because the strong allele for each of these lines generated very unhealthy plants. I also investigated the role of other proteins that were suspected to play a role in RDR6-RdDM. These mutant lines included the following proteins: DRM2, NERD, and SDE3. The SDE3 line was generated by using an artificial microRNA (amiRNA) to knock down expression because knocking out the SDE3 gene results in embryonic lethal plants. Influorescence is the tissue type utilized in all of the following experiments.

In order to generate the mutant lines, I began by planting individuals who were segregating for both *ddm1* and the mutation of interest. I first genotyped each individual for the mutation of interest. These mutations were usually created by the Salk Institute Genomic Analysis Laboratory, who insert a transfer-DNA (tDNA) sequence that contains specific primer-site targets into the gene of interest, thereby disrupting its sequence. The genotyping of these lines involved performing two rounds of PCR. One round used a primer specific to the insertion site in the genome and another primer specific to the primer-site in the tDNA insertion. If this PCR amplified, then the plant had an insertion into the gene of interest. The second round of PCR used primers specific to the area of the genome surrounding the insertion site. If this PCR amplified, then the plant did not have the insertion into the gene of interest,

because the insertion sequence was too large to amplify by conventional PCR. I was specifically looking for plants which were homozygous for the insertion sequence. Heterozygous and homozygous wild-type plants were discarded.

Once the plants which were homozygous mutant for the protein of interest were identified, I genotyped these plants for the *ddm1* mutation. The *ddm1* mutation is a single-nucleotide polymorphism which cannot be distinguished from wild-type DDM1 by PCR alone. This area must be amplified, and then digested by a restriction enzyme, which cuts at the wild-type site but not at the polymorphic site. The plants which were homozygous for the *ddm1* mutation were kept for further analysis.

I began my analysis of the role of various AGOs in the RDR6-RdDM pathway by performing qRT-PCR to measure levels of AGO mRNA accumulation in a *ddm1* versus Col background. If there were more of a particular AGO mRNA accumulating when TEs are active versus inactive, then this would suggest that this AGO was involved in the regulation of TEs, and a likely candidate to play a role in RDR6-RdDM. This required a set of primers specific to each AGO cDNA in order to amplify the proper target mRNA in each background. The mRNA of AGOs 1-10 were analyzed via qPCR in both the Col and *ddm1* backgrounds.

Next, I analyzed the levels of DNA methylation at the *Athila6* 5' promoter region (identified in Figure 7) to reveal which AGO proteins were necessary to initiate DNA methylation through RDR6-RdDM. This was done by bisulfite sequencing the *Athila6* 5' promoter region in each of the double mutant backgrounds, as well as in Col and *ddm1*. The results are reported as relative levels of CG, CHG, and CHH methylation. The mutant line *ddm1/ago10* was excluded from this analysis due to suspected contamination issues.

In addition to DNA methylation levels at the *Athila6* promoter, I also analyzed the levels of

Athila6 expression in Col, *ddm1*, and each one of the *ago/ddm1* mutant backgrounds using qRT-PCR. I used primers specific to the *Athila6 gag/pol* region of the cDNA in the qPCR analysis. The *gag/pol* region of the *Athila* transcripts is targeted because it is the region immediately downstream of the 5'LTR transcriptional start site. This way, the effect of start site methylation levels on *Athila* transcript levels can be observed. The levels of *Athila6* expression of the double mutants are compared to Columbia and *ddm1*. If there is a statistically significant increase in *Athila6* expression when compared to *ddm1* when the function of a particular AGO is lost, then it is likely that the protein plays an important role in silencing TEs in order to reduce the amount of *Athila6* mRNA accumulation.

In order to generate a clearer picture of the roles that each of the AGOs play with relation to the levels of various size classes of small RNAs (sRNAs), a small RNA Northern was performed to analyze the following backgrounds: Col, *ddm1*, *ddm1/ago1* weak allele, *ddm1/ago2*, *ddm1/ago4*, *ddm1/ago6*, *ddm1/ago7*, and *ddm1/ago9* strong and weak allele. The difference between the 24nt and the 21/22nt small RNA classes can be identified on this gel. Higher levels of 21/22nt sRNAs indicate that the RNAi pathway is functioning, while higher levels of 24nt sRNAs are generated by the PolIV-RdDM pathway.

After the identification of the particular AGO functioning in RDR6-RdDM was complete, I continued to explore other proteins potentially involved in RDR6-RdDM by looking at *Athila6* promoter methylation in *ddm1/drm2*, *ddm1/sde3*, and *ddm1/nerd* double mutants. Once again, the *Athila6* methylation levels were analyzed using bisulfite sequencing.

Finally, I explored the possibility that TEs could gain the ability to fully transpose when both RDR6-RdDM and PolIV-RdDM regulation are lost. I did this by performing a transposable element display, which targeted *Athila2* and *Athila6* in a *ddm1/polIV/rdr6* triple mutant background. The

transposable element profile of this triple mutant background was compared to Col. If any bands were present in *ddm1/polIV/rdr6* and not in Col, then it is likely that *Athila6* or *2* lost enough transcriptional regulation in the absence of both RDR6-RdDM and Pol4-RdDM to fully transpose, and this particular band would be the subject of further investigation in order to confirm this potential transposition.

Data and Analysis

The analysis of AGO mRNA expression levels did not show any significant differences between when TEs are active in a *ddm1* background versus when they are silent in a Col background (Figure 14). However, the quantitative PCR data of the *ddm1/ago* double mutants shows that when the function of AGO6 is lost, *Athila6* expression significantly increases when compared to *ddm1*, with a p-value equal to .0469 (Figure 15). The loss of AGO4 also generates an increase in *Athila* expression, albeit statistically insignificant, which is very close to those levels observed in a *ddm1/ago6*.

The bisulfite sequencing data shows that a loss of function of any of the AGO proteins results in a significant decrease in all types of methylation (Figure 16). However, the most significant decrease in methylation levels is observed when the functions of AGO1 or AGO6 are lost, indicating their important roles in the silencing of TEs. It also appears as though *ddm1/ago9* WA has a significant decrease in methylation levels. However, methylation levels of *ddm1/ago9* SA are not as significantly low as they are in *ddm1/ago1* or *ddm1/ago6,* and the strong allele is the full loss of function of AGO9. Therefore, it can be concluded that the function of AGO9 is not as critical to the methylation of TEs as the function of AGO6 or AGO1.

The small RNA levels determined by the sRNA Northern show that when AGO1 is lost, the productions of 21/22nt siRNAs are lost. There also appears to be a slight decrease in the levels of 21/22nt siRNAs in the ddm1/ago2 background, suggesting that it could also play some sort of role in

the production of these siRNAs. All other siRNA size classes persist in the other *ddm1/ago* mutant backgrounds (Figure 17).

Also, the analysis of the other potential components of RDR6-RdDM show that levels of methylation are lost when both NERD and SDE3 are not fully functional, but in particularly when DRM2 function is lost (Figure 18). This further proves that nearly all of the methylation seen in a *ddm1* background is *de novo*, because when the only *de novo* methyltransferase in Arabidopsis, DRM2, is lost, methylation levels decrease to almost non-existant levels.

The transposable element display set out to explore the possibility that when both RDR6-RdDM and Pol4-RdDM function are lost, TEs have the ability to transpose. No new bands, and therefore transpositions, were observed in *ddm1/rdr6/pol4* when compared to Col (Figure 19). The arrows on the gel indicate bands which are present in Col but not in *ddm1/rdr6/pol4.* It is likely due to the recombination of these particular elements out of the genome during the generation of the triple mutant line.

Conclusion

According to the qPCR and bisulfite sequencing data, AGO6 and AGO1 are the two most important components in the epigenetic silencing of TEs. Upon the analysis of the small RNA Northern data, it is evident that AGO1 is involved in the biogenesis of 21/22nt siRNAs which is functioning upstream of the *de novo* methylation process observed in RDR6-RdDM. There is no loss of any size class of siRNA when the function of AGO6 is lost, hinting towards the conclusion that it is the AGO responsible for incorporating siRNAs and using their sequence homology to target TEs for *de novo* silencing. Also, the bisulfite data from the other proteins of interest indicate that SDE3 and NERD are playing a role in the *de novo* methylation of TEs, but that DRM2 is really a key player in this pathway.

Combined with other data generated by our group, our latest model of RDR6-RdDM has it functioning alongside of Pol4-RdDM to complete the *de novo* silencing of TEs (Figure 20). AGO6 is the main AGO in the targeting of TEs for *de novo* methylation, potentially utilizing both 21/22 and 24nt siRNAs generated by Pol4-RdDM and RNAi. AGO4 has also been shown to be incorporating 24nt siRNAs to aid in the *de novo* methylation of TEs, but plays a less significant role in the RDR6-RdDM process than AGO6 (Figures 15 and 16). Since SDE3 is an RNA helicase, it is likely the helicase responsible for unwinding the double stranded RNA which is used to produce 21/22nt siRNAs [3]. The exact function of NERD remains unknown, but there is a possibility that it could be functioning at some point throughout the RDR6-RdDM pathway. Since DRM2 is the only *de novo* methyltransferase in Arabidopsis, it is likely that it is functioning downstream of both RDR6-RdDM and Pol4-RdDM at the intersection of the two pathways. If DRM2 is not present, then neither one of the pathways is able to lay down its *de novo* methylation on the TE targets.

The Role of RDR6-RdDM in the Resilencing of Transcriptionally Active Transposable Elements.

Materials and Methods

To begin to explore the potential role of RDR6-RdDM in the transcriptional resilencing of TEs, I began by setting up a cross which is depicted in Figure 21. I crossed one parent which was homozygous mutant for the protein of interest and homozygous wild type for DDM1 with another parent which was homozygous mutant for the protein of interest and homozygous mutant for DDM1. This type of approach is known as a "corrective reestablishment of silencing assay". This results in offspring which are homozygous mutant for the protein of interest and heterozygous for *ddm1*. In the positive control, Columbia crossed to a *ddm1* mutant, all of the methylation and expression of *Athila6*

elements is restored to wild type levels. It is known that Pol IV-RdDM is involved in this transgenerational resilencing of TEs [4]. If RDR6-RdDM contributes to this resilencing as well, then when the function of its components are lost, the full restoration of methylation on *Athila6* will not occur. The RDR6-RdDM components of interest analyzed in these experiments include RDR6 and DCL2. Pol IV-RdDM components were explored as well, and the resulting affects on *Athila6* were compared to those that occur when components of RDR6-RdDM were lost.

I first analyzed the levels of methylation at the promoter of *Athila6* in each one of the aforementioned backgrounds. This was done using bisulfite sequencing targeting the 5'promoter region of *Athila6*. I also analyzed the levels of *Athila6* expression in these backgrounds using qRT-PCR. The methylation and expression levels of these mutants were compared to Col and *ddm1* heterozygous lines to analyze the importance of each of these components in the corrective resilencing of TEs.

Data and Analysis

The data generated from the bisulfite sequencing of the mutant lines show that wild-type levels of methylation cannot be properly restored when components of RDR6-RdDM, such as RDR6 and DCL2, are non-functional (Figure 22). Levels of *de novo* methylation increase and maintenance methylation decrease in these particular mutant backgrounds when compared to Col and *ddm1* het. The qPCR data from these backgrounds shows that when the components of either RDR6-RdDM or Pol4-RdDM are lost, expression of TEs significantly increases when compared to the *ddm1* het background where resilencing is properly occurring (Figure 23). The methylation levels are lower and the *Athila* expression levels are higher in Pol4-RdDM mutants than they are in RDR6-RdDM. These trends hint towards the conclusion that Pol4-RdDM generally seems to be playing a larger role than RDR6-RdDM in the resilencing of TEs.

Conclusion

Levels of *de novo* methylation must be reestablished after every cellular division when components of RDR6-RdDM or Pol4-RdDM are lost because maintenance methylation cannot be properly established in these backgrounds. Our current model shows that Pol4-RdDM must be occurring on the methylated TEs inherited from the DDM1 wild-type parent while RDR6-RdDM is occurring on the unmethylated TEs inherited from the *ddm1* parent in order to properly reestablish methylation on the unmethylated TEs (Figure 24). Once again, 24nt siRNAs are being generated by the Pol4-RdDM pathway, with are incorporated into both AGO6 and AGO4. Simultaneously, 21/22nt siRNAs are being generated by RDR6-RdDM, and are also being incorporated into AGO6. Together, the AGOs are recruited to a scaffolding transcript produced by Pol5, and recruit the methyltransferase DRM2 to fully remethylate active TEs. The fact that Pol5 is at the intersection of these two transcriptional silencing pathways explains the dramatic loss of methylation and gain of TE expression when Pol5 activity is lost (Figures 22 and 23).

The Potential of RDR6-RdDM to Resilence TEs Beyond *Athila*

Materials and Methods

The biological relevance of the resilencing of TEs by RDR6-RdDM would be increased if it were shown that this pathway works to regulate TEs other than *Athila6*. Another TE that our lab was interested in investigating is ENSPM5, which is a DNA element in the Arabidopsis genome. We were interested in investigating ENSPM5, because when RDR6 activity was lost, the levels of 21/22nt siRNAs (indicative of RDR6-RdDM activity) which matched the ENSPM5 region decreased, while levels of 24nt siRNAs (indicative of Pol4-RdDM activity) remained the same when compared to a

ddm1 background [14]. Therefore, if an increase in ENSPM5 expression levels is observed in RDR6-RdDM mutant backgrounds, then it can be directly attributed to a reduction in RDR6-RdDM activity. In order to investigate the expression levels of ENSPM5, I designed primers specific to the cDNA of this DNA element and performed qPCR. The backgrounds analyzed in this experiment include the following: Col, *ddm1*, *+/ddm1*, *+/ddm1*; *rdr6*, *+/ddm1*; *dcl2*, *+/ddm1*; *dcl4*, *+/ddm1*; *rdr2*, *+/ddm1*; *dcl3*, *+/ddm1*; *pol4*, *+/ddm1*; *pol5*, where *+/ddm1* indicates a *ddm1* heterozygote. The results of this set of qPCR analysis was compared to the qPCR analysis of *Athila6* to see if there was a similar pattern of TE expression.

Data and Analysis

In each one of the mutant backgrounds, *ENSPM5* shows similar levels of expression to *Athila*. *ENSPM5* levels of expression increase as expected when Pol4-RdDM components are lost, but also when RDR6-RdDM components are lost as well (Figure 25). Once again, we see Pol4-RdDM playing a larger role in the resilencing of expression than RDR6-RdDM, just as with *Athila*.

Conclusion

RDR6-RdDM functions more ubiquitously than exclusively working to silence *Athila6*. The data from this experiment shows that when its components are lost, the expression of ENSPM5 significantly increases when compared to *ddm1* background levels. Since *ENSPM5* is a DNA element, this shows that RDR6-RdDM can function to regulate a wide array of TEs in the genome. Future work will be to identify other targets of this pathway in the genome to understand its importance in the maintenance of genomic integrity.

The Time Point in the Development of Arabidopsis in which the Resilencing of TEs Occurs

Materials and Methods

After the confirmation of the involvement of RDR6-RdDM in the transgenerational resilencing of TEs, I set out to identify the time point in the development of Arabidposis during which the resilencing occurs. If this time point could be pinpointed, then this would serve as the ideal developmental stage in which to study the mechanism of RDR6-RdDM. The point in development when *Athila* expression is insignificantly different between *ddm1* heterozygous; *rdr6* homozygous plants and *ddm1* heterozygous plants is the time before RDR6-RdDM begins working to shut down TE expression, and would be the interesting time point to focus on for future analysis. Depending on the availability of tissue, some plants which were mutant for either components of RDR6-RdDM or PolIV-RdDM were analyzed as well to look for any interesting trends.

The analysis of one of the most mature tissues in the plant, the inflorescence, had already been completed during the investigation of the resilencing of TEs through RDR6-RdDM (Figure 23). I moved earlier in developmental time to the juvenile leaf tissue in order to perform qRT-PCR of *Athila6* transcripts in the following backgrounds: +/*ddm1;rdr6, +/ddm1;dcl2, +/ddm1;polIV* , and +/*ddm1*. I then moved back in developmental time to investigate Athila6 transcript levels in both seedling and seed tissue in the following backgrounds: Col, *ddm1*, +/*ddm1*, and +/*ddm1;rdr6*. Seed tissue is the youngest tissue analyzed in this particular set of experiments because younger tissue types were not saved.

Data and Analysis

Quantitative PCR data from all tissues types shows that *ddm1* het levels of expression remain lower than the *ddm1* het; *rdr6* mutants throughout the development time points investigated (Figure 26). However, the gap in expression levels between the two backgrounds does get more narrow as the

tissues move earlier in developmental time.

Conclusion

While I did not pinpoint the exact developmental time point in which the resilencing of TEs occurs, I did show that *Athila* expression between +/*ddm1* and +/*ddm1*;*rdr6* does become more similar in younger tissues, such as seedling and seed. I also concluded that the resilencing in the WT Col x *ddm1* individuals occurs very early in development before the mature seed is formed. The future direction of this project will be to go back to an earlier tissue type than seed such as embryonic tissue in order to determine if the resilencing of TEs is a gradual process, or if it occurs right after the moment of fertilization.

Final Conclusions

Before I contributed to the research in my group, little was known about our lab's primary TE, *Athila*, and the RDR6-RdDM pathway was still only beginning to be understood. The common dogma in the field was that Pol4-RdDM was the sole pathway functioning to transcriptionally silence TEs in the Arabidopsis genome, and that the only role of the components of RNAi was to post-transcriptionally represses active TEs. Throughout my time as a researcher in the Slotkin lab, I have increased our general knowledge of the molecular characteristics of *Athila*, as well as contributed to understanding the mechanism of the newly-identified RDR6-RdDM pathway. Because of my work, more is understood about this important pathway which works to prevent dangerous genomic damage and instability by shutting down TE activity.

This research could have multiple important future implications. For one, the multi-billion dollar

industry of the generation of genetically modified organisms (GMOs) could benefit from such research. Currently, scientists insert transgenes of endogenous genes that they would like to silence via RNAi. Unfortunately, this method of gene silencing is not heritable, and it becomes costly and inefficient to insert transgenes into each new generation. Understanding how RNAi helps to trigger the initiation of transcriptional silencing could lead scientists to develop heritable changes in selective gene silencing. The initial transgene could then be segregated out of the line so that the genome no longer contains any foreign DNA. This could alleviate many of the complications associated with the generation of transgenic organisms. In addition, TE activity is the cause behind several human diseases, such as hemophilia A and B, severe combined immunodeficiency, and Duchenne muscular dystrophy [7] If TE activity could be silenced using a pathway similar to RDR6-RdDM, many of these devastating health disorders could be avoided.

Acknowledgments

I would like to thank the Slotkin lab for their continued support of my research endeavors. I would also like to thank my funding sources, which include the National Science Foundation Research Experience for Undergraduates Awarded as a Supplment to the Slotkin Grant MCB-1020499, the Arts and Sciences Undergraduate Research Scholarship, the Elizabeth Wagner Scholarship through the Department of Molecular Genetics, and the Summer Undergraduate Research Fellowship through the American Society of Plant Biology.

Works Cited

1) Belancio, V. P., D. J. Hedges, and P. Deininger. "Mammalian Non-LTR Retrotransposons: For Better or Worse, in Sickness and in Health." *Genome Research* 18.3 (2008): 343-58. Print.

2) Dalmay, T. "SDE3 Encodes an RNA Helicase Required for Post-transcriptional Gene Silencing in Arabidopsis." *The EMBO Journal* 20.8 (2001): 2069-078. Print.

3) Garcia, Damien, Shahinez Garcia, Dominique Pontier, Antonin Marchais, Jean Pierre Renou, Thierry Lagrange, and Olivier Voinnet. "Ago Hook and RNA Helicase Motifs Underpin Dual Roles for SDE3 in Antiviral Defense and Silencing of Nonconserved Intergenic Regions." *Molecular Cell* 48.1 (2012): 109-20. Print.

4) Greenberg, Maxim V.c., Israel Ausin, Simon W.l. Chan, Shawn J. Cokus, Josh T. Cuperus, Suhua Feng, Julie A. Law, Carolyn Chu, Matteo Pellegrini, James C. Carrington, and Steven E. Jacobsen. "Identification of Genes Required for De Novo DNA Methylation in Arabidopsis." *Epigenetics* 6.3 (2011): 344-54. Print.

5) Gruntman, Eyal, Yijun Qi, R. Keith Slotkin, Ted Roeder, Robert A. Martienssen, and Ravi Sachidanandam. "Kismeth: Analyzer of Plant Methylation States through Bisulfite Sequencing." *BMC Bioinformatics* 9.1 (2008): 371. Print.

6) Jeddeloh, Jeffrey A., Trevor L. Stokes and Eric J. Richards. "Maintenance of genomic methylation requires a SWI2/SNF2-like protein." *Nature Genetics* 22 (1999): 94-97.

7) Kapitonov V.V., Pavlicek A., Jurka J., 2006, "Anthology of Human Repetitive DNA". Encyclopedia of Molecular Cell Biology and Molecular Medicine.

8) Kato, Masaomi, Asuka Miura, Judith Bender, Steven E. Jacobsen, and Tetsuji Kakutani. "Role of CG and Non-CG Methylation in Immobilization of Transposons in Arabidopsis." *Current Biology* 13.5 (2003): 421-26. Print.

9) Kim, J. K., M. Samaranayake, and S. Pradhan. "Epigenetic Mechanisms in Mammals." *Cellular and Molecular Life Sciences* 66.4 (2009): 596-612. Print.

10) Klenov, M. S., and V. A. Gvozdev. "Heterochromatin Formation: Role of Short RNAs and DNA Methylation." *Biochemistry (Moscow)* 70.11 (2005): 1187-198. Print.

11) Lander, E.S., Linton, L.M., Birren, B., et al (February 15, 2001). Initial sequencing and analysis of the human genome. *Nature London*-, 6822, 860-921.

12) Nuthikattu, S., A. D. Mccue, K. Panda, D. Fultz, C. Defraia, E. N. Thomas, and R. K. Slotkin. "The Initiation of Epigenetic Silencing of Active Transposable Elements Is Triggered by RDR6 and 21-22 Nucleotide Small Interfering RNAs." *Plant Physiology* 162.1 (2013): 116-31. Print.

13) Pontier, Dominique, Claire Picart, François Roudier, Damien Garcia, Sylvie Lahmy, Jacinthe Azevedo, Emilie Alart, Michèle Laudié, Wojciech M. Karlowski, Richard Cooke, Vincent Colot, Olivier Voinnet, and Thierry Lagrange. "NERD, a Plant-Specific GW Protein, Defines an Additional RNAi-Dependent Chromatin-Based Pathway in Arabidopsis." *Molecular Cell* 48.1 (2012): 121-32. Print.

14) Rij, Ronald P. Van, and Eugene Berezikov. "Small RNAs and the Control of Transposons and Viruses in Drosophila." *Trends in Microbiology* 17.4 (2009): 163-71. Print.

15) Slotkin, R. Keith. "The Epigenetic Control of the Athila Family of Retrotransposons in Arabidopsis." *Epigenetics* 5.6 (2010): 483-90. Print.

16) Vanyushin, Boris F., and Mikhail D. Kirnos. "DNA Methylation in Plants." *Gene* 74.1 (1988): 117-21. Print.

17) Wu, L., Mao, L., & Qi, Y. (July 30, 2012). Roles of DICER-LIKE and ARGONAUTE proteins in TAS-derived siRNAs triggered DNA methylation. *Plant Physiology*.

Figures for Characterizing the Novel Roles of Small RNA-directed DNA Methylation in the Epigenetic Regulation of *Athila* Family Retrotransposons

By
Erica Thomas
The Ohio State University
April 2014
Project Advisor: Dr. R. Keith Slotkin, Department of Molecular Genetics

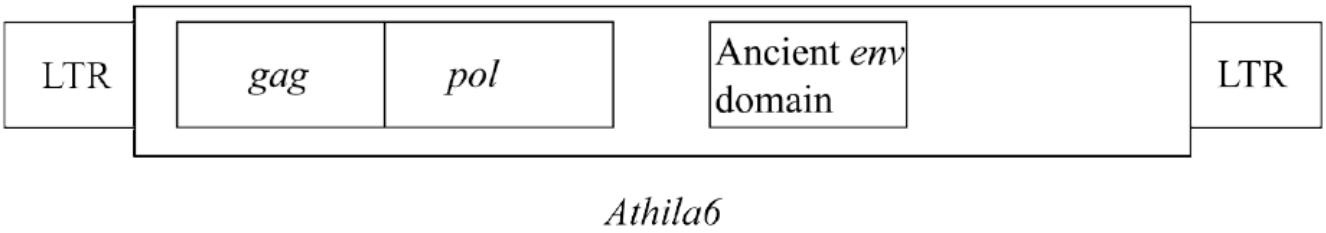


Figure 1: The structure of the *Athila6* LTR retrotransposon. Long Terminal Repeats (LTRs) flank either side of the element. The *gag* and *pol* genes within the element code for its transposition machinery. The ancient *env* gene is the nonfunctional domain of an ancient retroviruses envelope coding region.

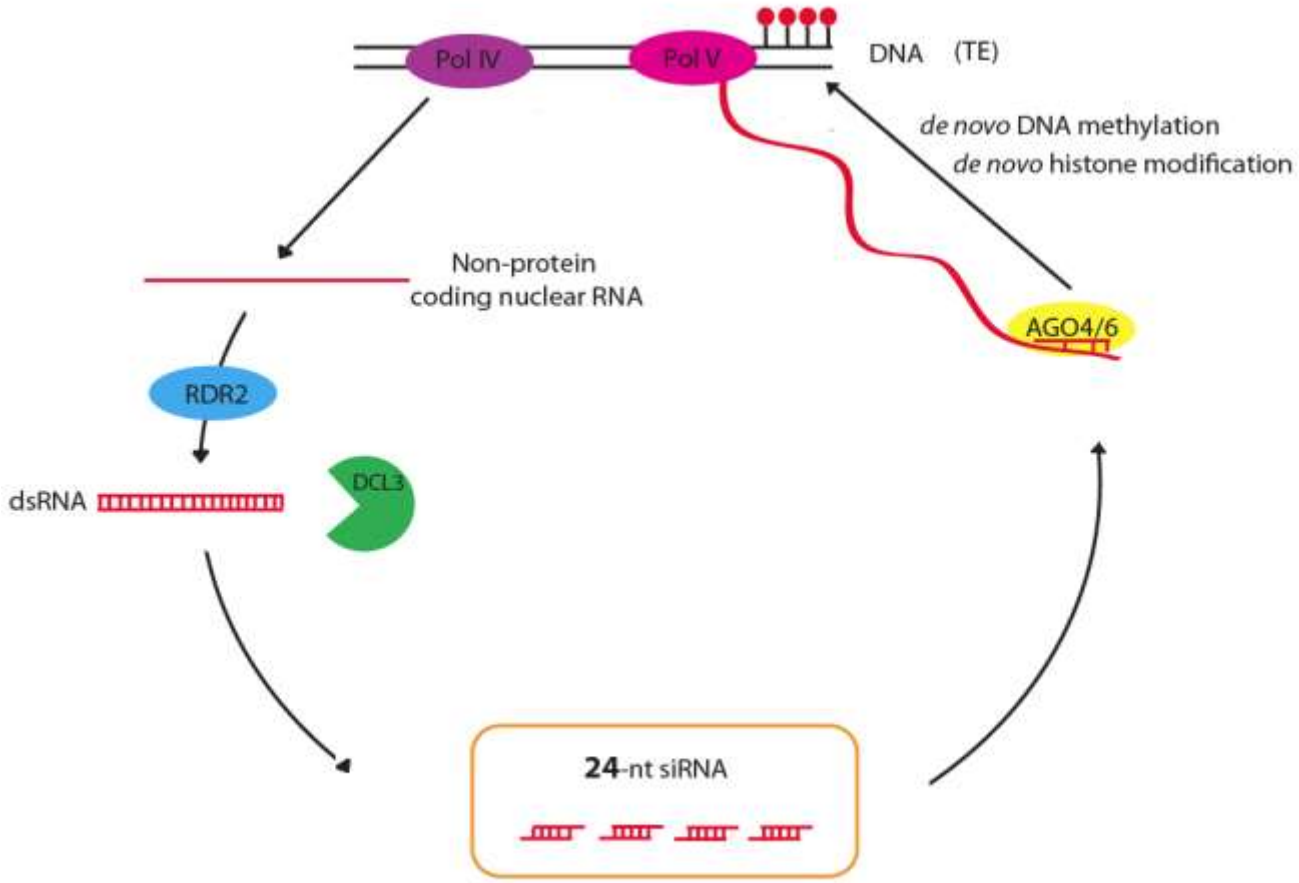


Figure 2: A diagram of the transcriptional gene silencing pathway known as RNA-directed DNA Methylation (RdDM). The components of this pathway work to silence TEs through the processing of non-coding transcripts generated by PolIV into 24-nt siRNAs. These siRNAs are then used to target TEs via sequence homology for symmetrical *de novo* DNA methylation and histone tail modification.

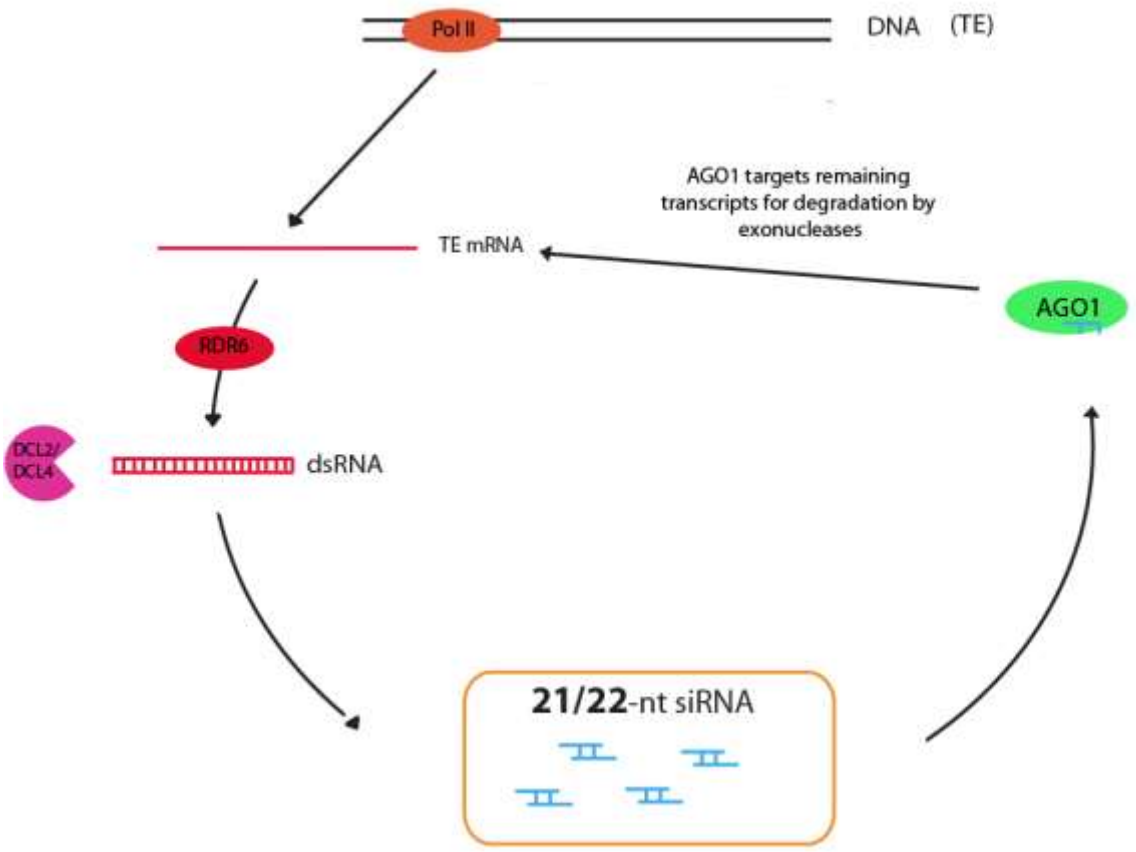


Figure 3: A diagram of the post-transcriptional gene-silencing pathway known as endogenous RNA interference (RNAi). The pathway begins with the selective targeting of TE mRNA by RDR6, which generates double stranded RNA. DCL2 and 4 then process this RNA into 21/22-nt siRNAs, and their sequence homology is then used to target remaining TE mRNA transcripts for degradation by exonucleases. This way, the protein products that are needed for TE transposition cannot be generated. Pol II can only be targeted to TE promoters when there is a global loss of methylation, which can be achieved through the loss of function of a protein known as Decrease in DNA Methylation 1 (DDM1).

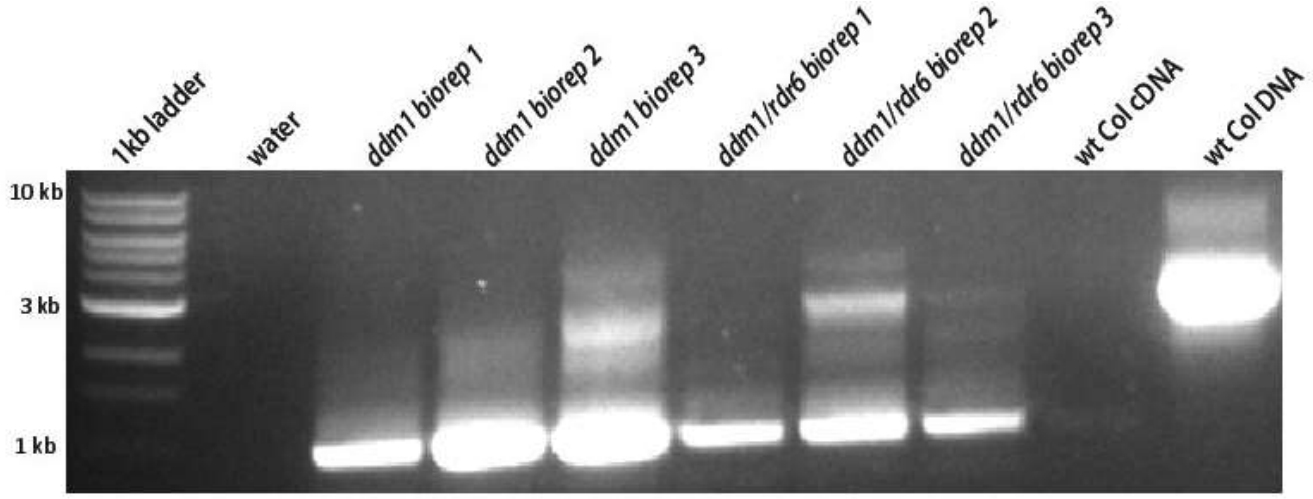


Figure 4: PCR products of *ddm1*, *ddm1/rdr6*, and Col cDNA as well as Col DNA run on an agarose gel using *Athila6* specific primers. The Col DNA resulted in a 3 kilobase pair (kb) band, as indicated by the blue arrow. The *ddm1* and *ddm1/rdr6* products were either around 1kb or 3kb, suggesting that there exist both spliced and unspliced *Athila6* transcripts when TEs are active. This assay also indicated that the intron is approximately 2kb in length.

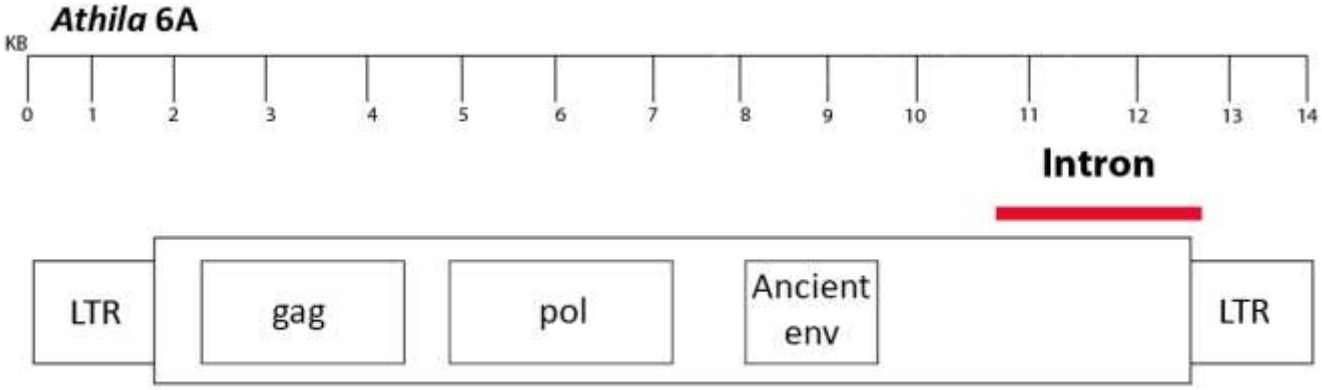


Figure 5: Final position of intron as determined by the sequencing of Athila cDNA. The intron resides in the 3'region of the *Athila* sequence, and is approximately 1,810 base pairs in length.

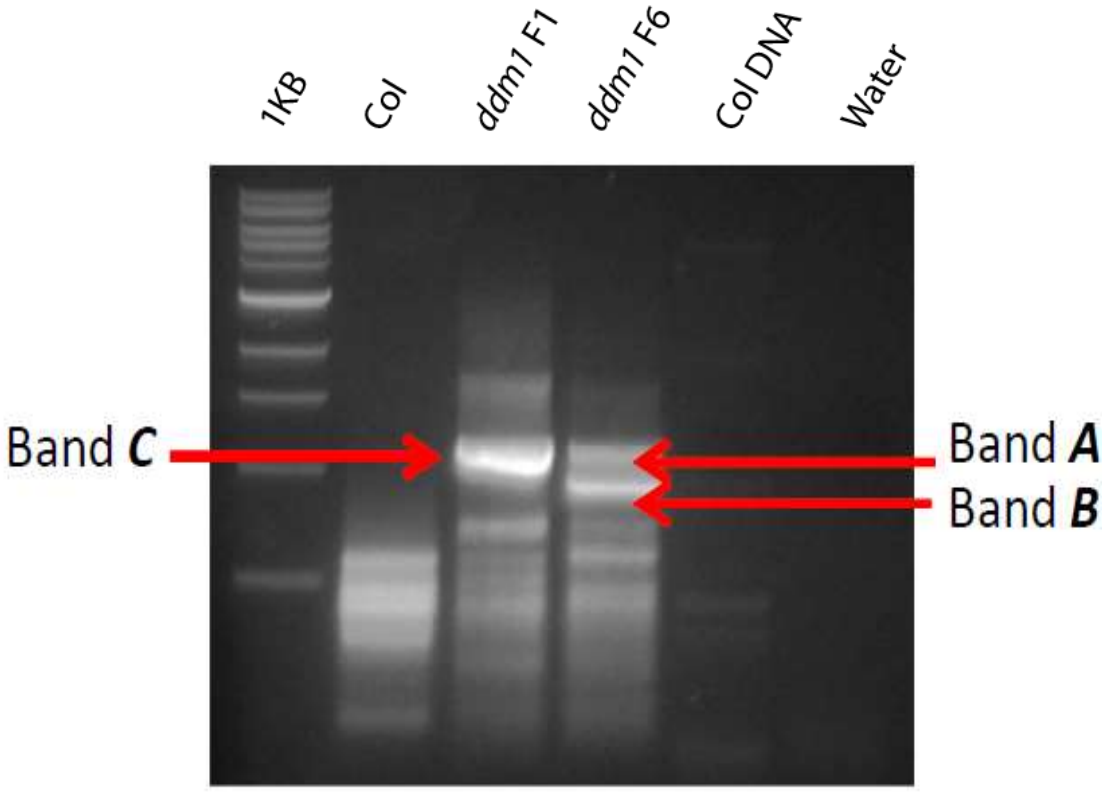


Figure 6: Results of 5'RACE of the start site of *Athila* transcription in its 5'LTR region. SL001 (1st generation *ddm1*) shows one distinct band, labeled as Band C. SL300 (6th generation *ddm1*) shows two distinct bands, labeled as Bands A and B. Each one of the labeled bands was gel extracted, cloned, and then sequenced to determine the exact location of the start site.

Athila6A LTR consensus:

TGATATGTCTATAATTTGCATGTTTTCAGTGTCCATTCATCATCGTTTTGAGTCCAGTTTCG
TATCATTCATCACTGTTTTATATCATTTCTCATCATTCTTGCATACTTTGCATGATTAGGAT
AGCTTTGCATACATATTGCATTTCTGAGTTGTTTTCAGGTGATTTGGAGCTGTTTGCGAGCA
AATTGGAAGAAACGAGCCAGAACCAGAAGACATACTCGACCCCCAGGTCGAGTGACTTTG
GAGCCATTCTTCCCACATACTCGACCCCCAGGTCGAGTGACTTTGGAGCCATTCTTCCCAT
CCACTCGACCACCAGGTCGAGTAACCTCAGCTCAGGCCACTCGATGACATTACTCGACCC
CCTGGTCGAGTATCACTTCGCCACACCACCTGACCACACTCGACCATTCACTCTACCACGT
TACTCGACCCCCTGGTCGAGTATCATCACTCACCACCATCACCATCACTCGACCGGACACT
CGATCACGTCTTCACAGTCTACTCAAATCCGCACTCAACCAGACAAGCTGAGCACAAGGA
AGAGAAGAGGAGAAGACAAAGTGTTTGGAAGCGGCCTGGACCTCCATCGGATCACGAAG
CCCATCTCGGCCCATTATCACTCTATGGGCCGGGCGATTAGGTTATTGGCCCGTCTACTAT
CATTTTATTTCGTTTTGTATAAATAGATGTCTTAGGGTTCTGTCCTGAGACXXATXCTXXXX
XXXAGXTCXXXGACATTGAXGTTTTTGCTTCAGTTTTATTTTCTGTTTTACTCTGCTGCGCC
GCTTTTGCTTCTGCAACCTGTAATTCGAGATTTTTCCAAGTTATTCAGATTCCGCNTTTGAT
TTCATCTGTCATCTTGTTTCTCTACTCTTTATCTCTTTACTTATGCAATATTATCGTTTATCT
GCGTTTATGTCTTTXXXXGAXGXCATGTTGTCTGAGTAGTGACTTAGAATTCTTAGGGATG
GGATAGAGTAGTTGTGGAATCCGTAGTCTGTAGAATGGTTAAGTTTTAGAATTGATTGAAT
CCCTTTAGGACTAGTTGCGTTTACTGCTTATTTCTTTCTGATCAACTGGAATTCGATCCCAA
GCATTCCCGCACCCAGAAGGTGTTCGATGGAATGCTTGATCCACTAGTTCCTGAGATATGC
GTCTCTATCCCAAGGGATTGGCCGTTTAGAGCGTTTATTGACTTTATCAGTCTGTTCTTATT
GCCTGCATAGTTAGATTCCGTTAATGGGAATTAGTCTGGGCTAGCTTTGCTTGAGGATTTC
TATCACGGGAGTGAATTGATCTGTTGTTGAACCTGTTGTCTAGGGATAGCTTGATTGCGCT
TGTTAACCATTCGAATAGGCTAGGATACCACTCTACTCGATTACCCCATCCTTAGGAATTT
CTCGTCTATCTTATTTCTCTGTTTTATCGTTATCGCTTAATCGTTCTCATTGCCTGTTTCTTA
GTTTTTACAATCACTCGACCAACTTACTCGACCACACCCAGTGTCTGGCAACAGACTGTGC
AGTCGAGTANNGTTGTCTTGTTTCCTGTCTGTTACTCGACCAGTATACTCGACCACACCCA
GTGTCTGGCAACAGACTGTGCAGTCGAGTATAGCTGTTTCATATCCTGTCTGTTACTCGAC
CTGTTACTCGACCACATCCTACTTCTGGCATCAGCCTGTTGTGGTCGAGTGATTTTAGTAAT
TCTGTTATTGCTTCTGTTTTCTGCATGTTCGCTTAGGACTGTTAGAAACCCCAAAACTGTTA
TTGCTTGGCTTGACTTAGTGACTTCTGATCACATCTCATCTGTTTGCATCACACCCATTTGG
ATTGACACCTAAAATACTACAACGACATGATTGGTGTTTTAGGAATAATTGACTAAAAAC
CTATTATCA

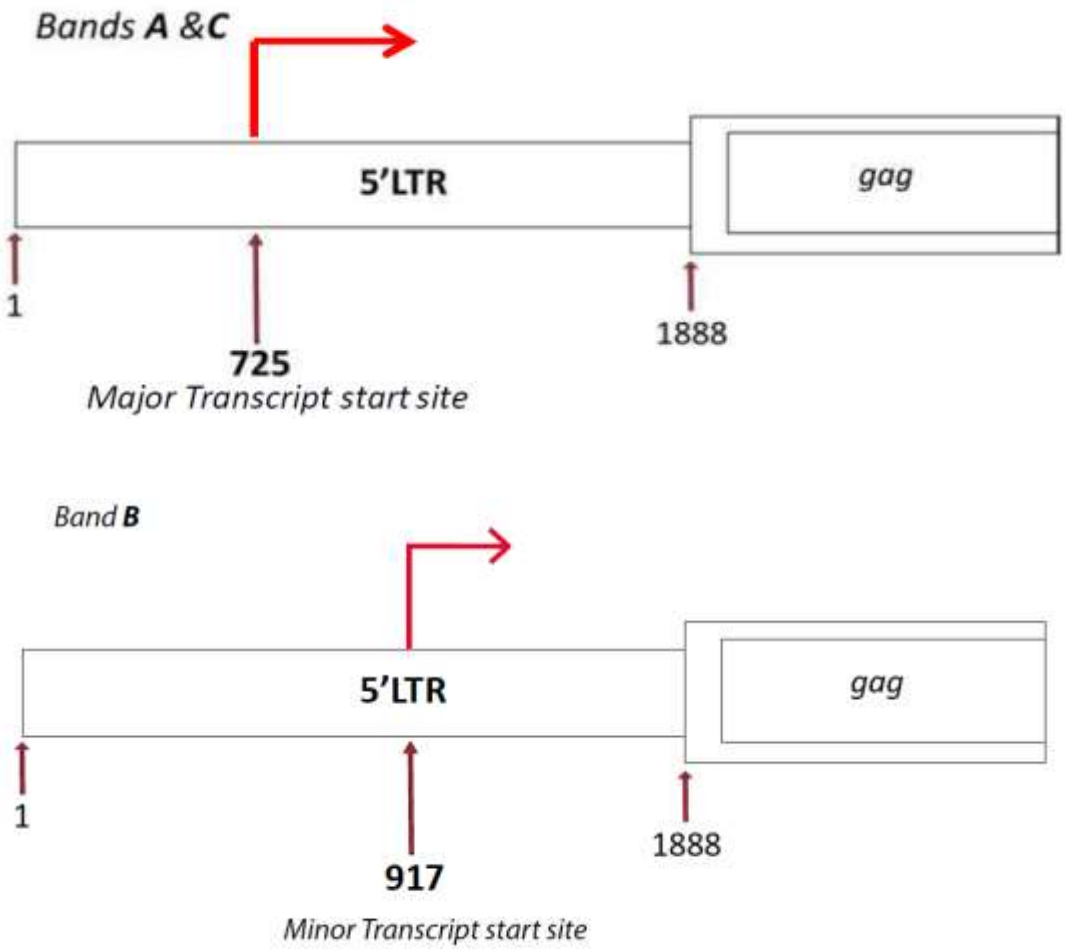


Figure 7: Major and minor transcriptional start sites of *Athila6*. The sequencing of bands A from SL300 and C from SL001 produced a start site that begins at base pair 725 of the *Athila* LTR consensus sequence. The sequencing of band B from SL300 produced a start site that begins at base pair 917 of the *Athila* LTR consensus sequence. This second start site is considered a minor start site because only one specific *Athila* element in the Arabidopsis genome (At1g40095) exhibits this particular start site.

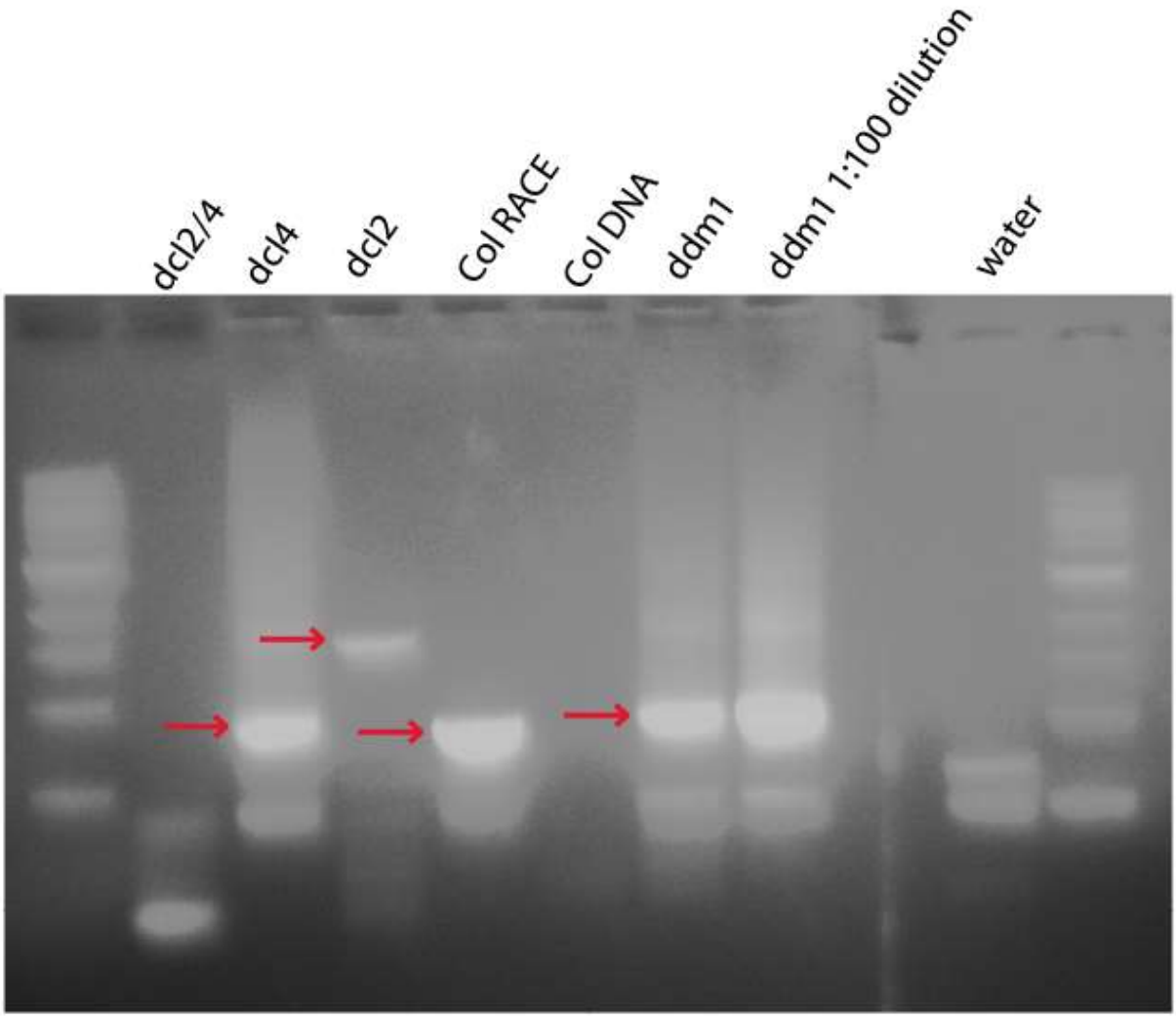


Figure 8: Results of 5'RACE of the start site of *Athila* transcripts produced in the *env* region. The most distinct band in *dcl4*, Col RACE, *ddm1*, and *ddm1* 1:100 dilution were gel extracted and sequenced. The larger band in *dcl2* was also gel extracted and sequenced in order to identify the exact location of its start site. During 5'RACE, samples undergo many rounds of PCR, which results in the amplification of even small amounts of transcripts. Since *Athila* expression is observed in the Col RACE sample, it is likely that there is a small amount of TE expression in the WT background.

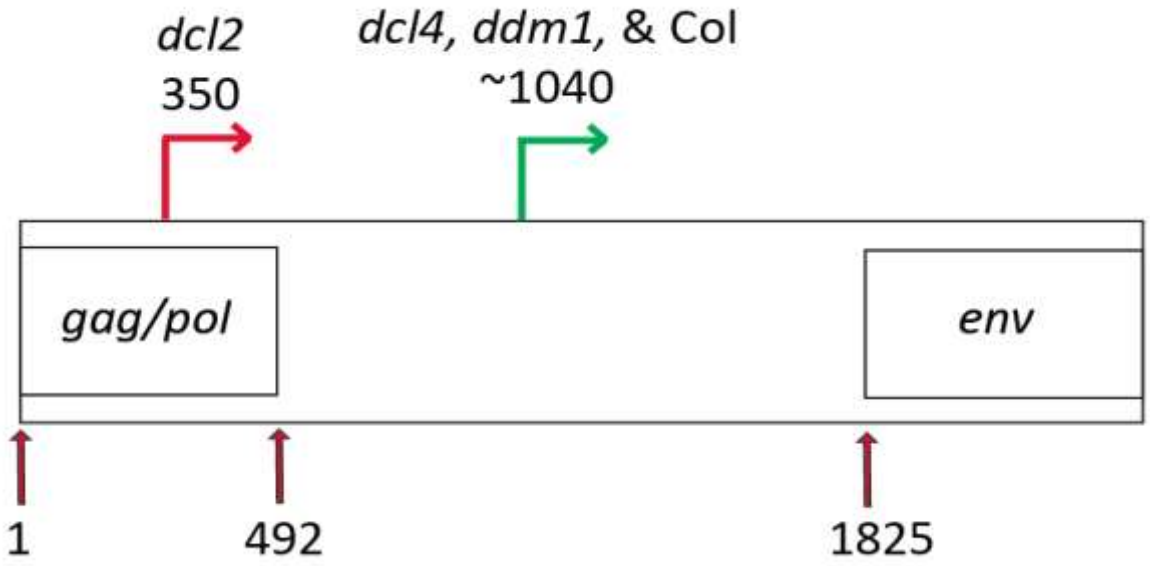


Figure 9: The region of interest in this particular figure shows the last 492 base pairs of the *gag/pol* region in *Athila*, up to the beginning of the *env* region, which is approximately 1825 base pairs later. The start site of the transcripts produced in *dcl2* is around base pair 350 of the last 492 base pairs of *gag/pol*, while the start sites of *dcl4*, *ddm1*, and Col begin at around base pair 1040 in the intergenic region.

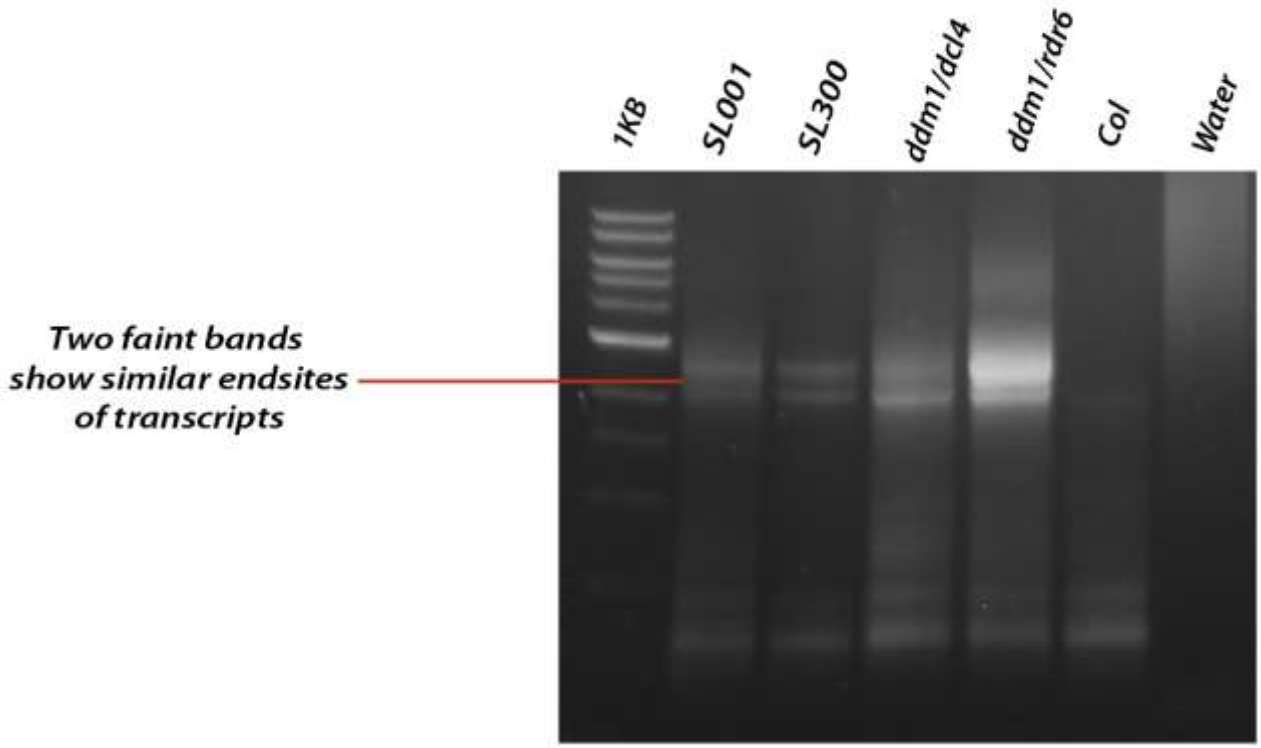


Figure 10: Results of 3'RACE of the end site of *Athila* transcripts produced in the 3'LTR region. As the label indicates, two faint bands were observed in each one of the backgrounds excluding Col. Each one of the bands was gel extracted, cloned and sequenced to identify the exact stop site of *Athila* transcription.

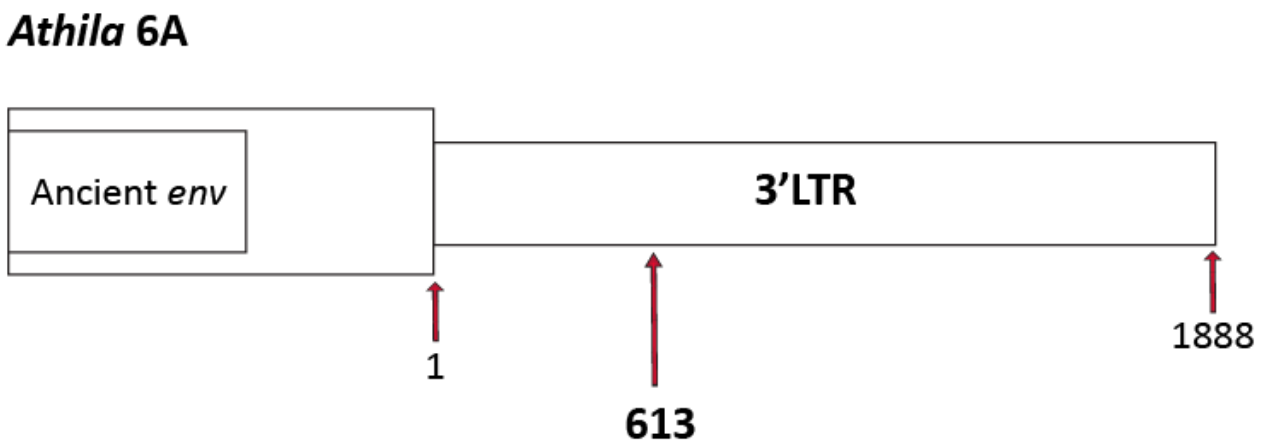


Figure 11: Diagram indicating the stop site of *Athila* transcription. The sequencing results of the gel-extracted bands from 3'RACE show an end site around base pair 613 of the 3'LTR *Athila* consensus sequence.

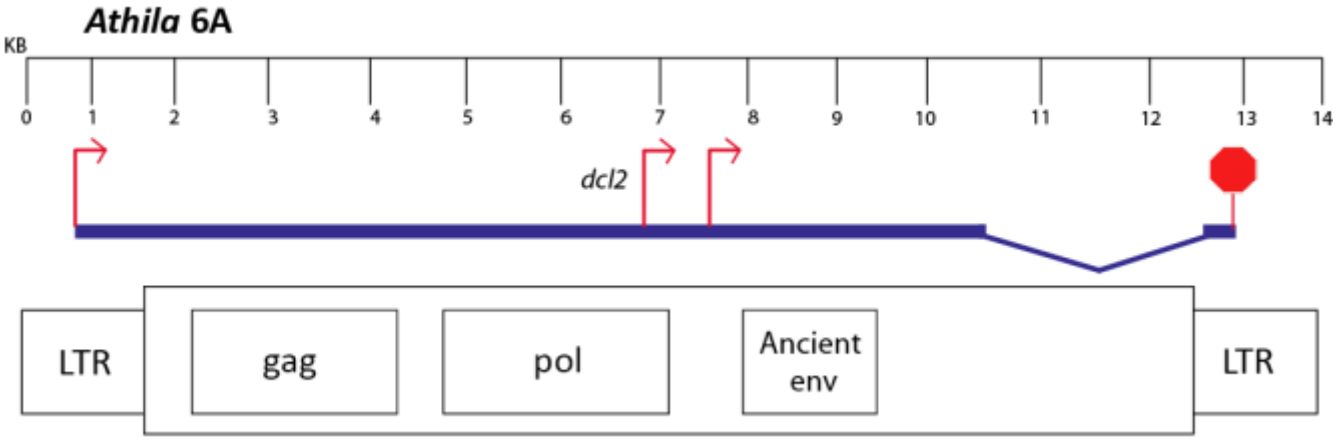
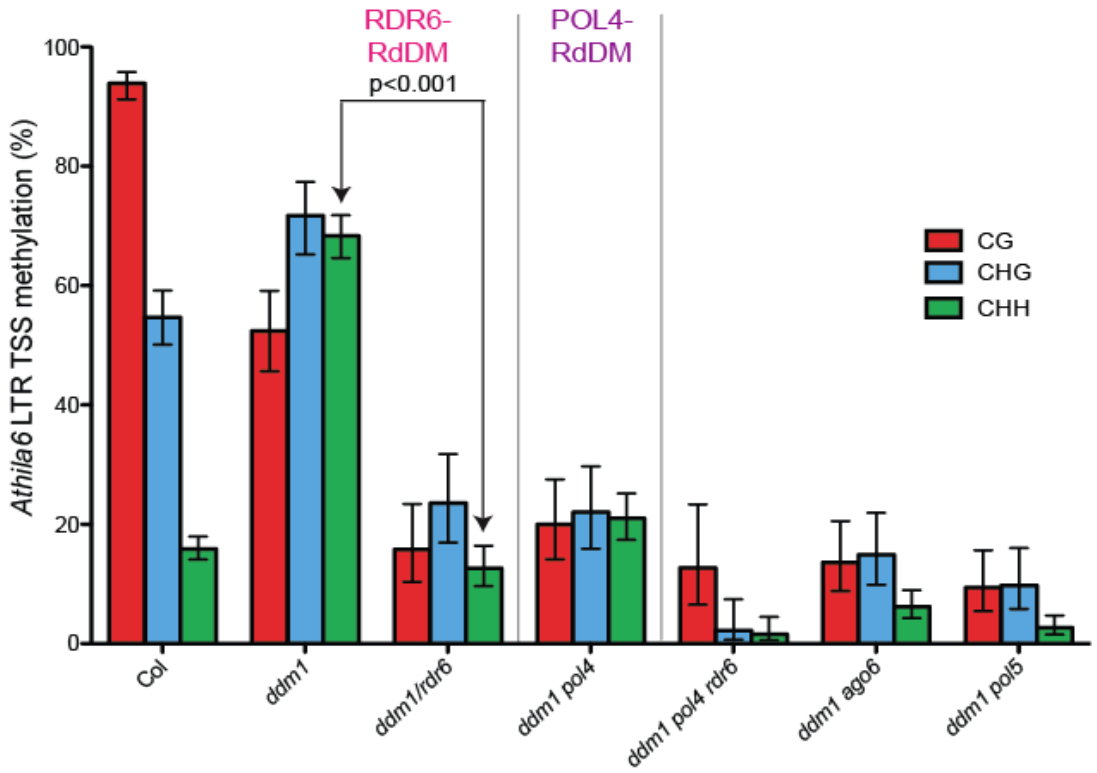


Figure 12: The complete model of the molecular characterization of *Athila6*. The main transcriptional start site in the 5'LTR is depicted, as well as the *dcl2*-specific start site at the end of the *gag/pol* region and the main *env* start site found in the intergenic region. The intron is depicted on the 3' side of the element, as well as the end site located in the 3'LTR. Future directions of this research include investigating if the full length *env* transcript is spliced, identifying the *gag/pol* transcripts which remain unspliced, and pinpointing the difference between the *env* and *gag/pol* transcriptional stop sites.

A)



B)

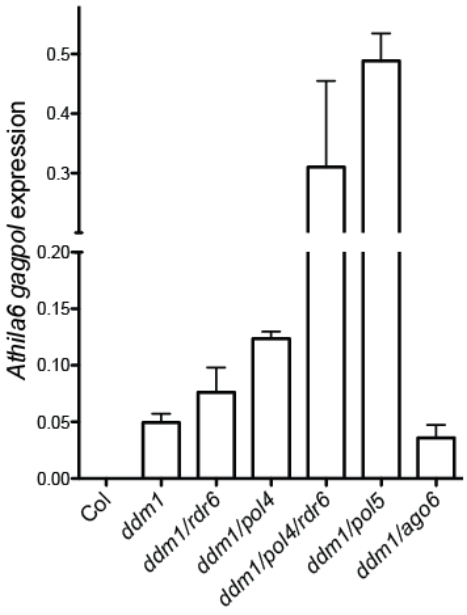


Figure 13: Previous data generated by our group shows that both Pol4-RdDM and RDR6-RdDM are important for the epigenetic silencing of active transposable elements. Previously, it was understood that proteins such as RDR6 and AGO6 only function in endogenous RNA interference, but the bisulfite data (A) shows that DNA methylation is lost when the function of these components is compromised. This observation is also supported by the quantitative PCR data (B) showing increased *Athila6* expression when RDR6 or AGO6 are lost. Our group termed this methylation by components of RNAi as RDR6-RdDM.

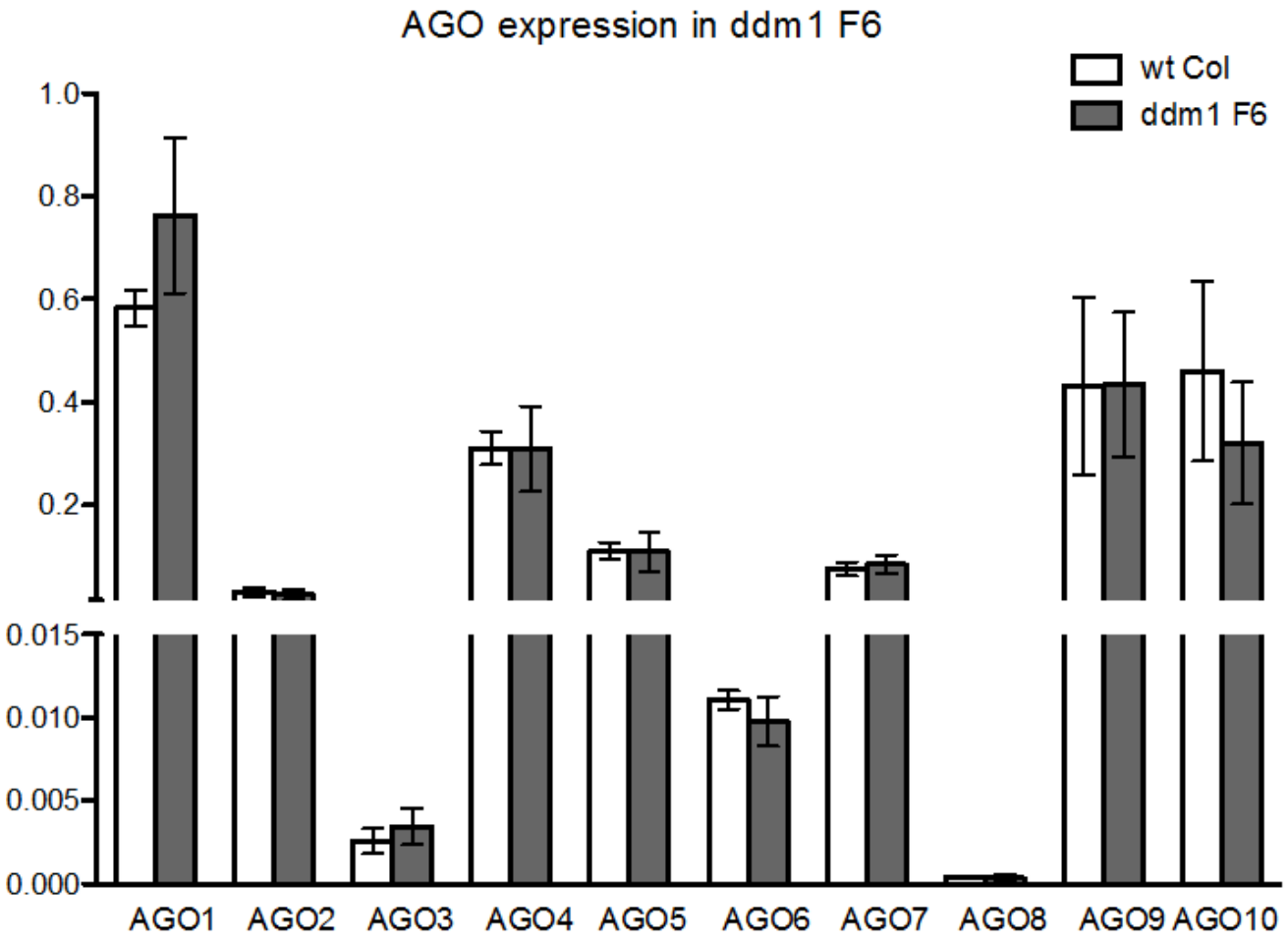


Figure 14: Quantitative PCR targeting various AGO mRNA levels in a *ddm1* versus Col background. There is no significant upregulation of any AGO mRNA expression in a *ddm1* background when TEs are active compared to when TEs are silent in a Col background.

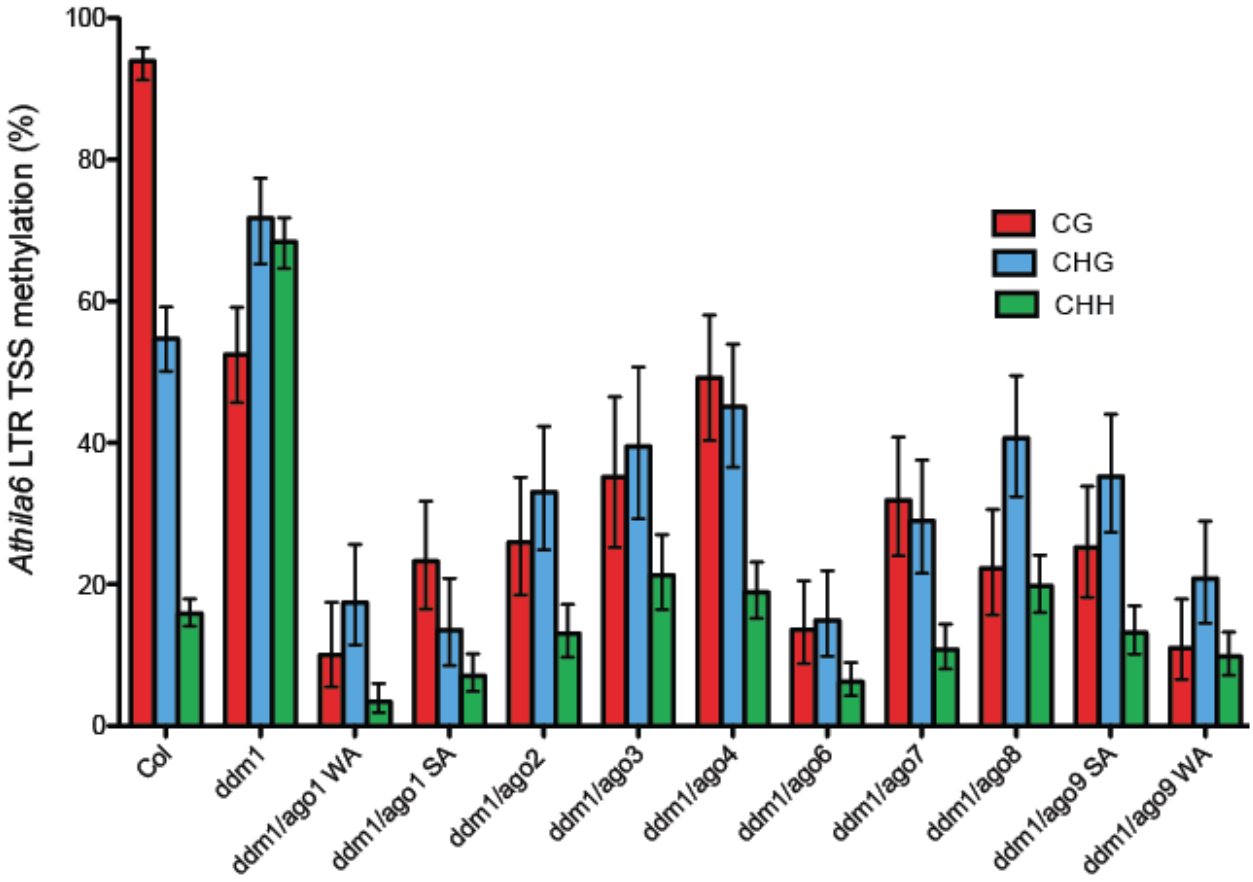


Figure 15: Bisulfite sequencing data from various *ddm1/ago* mutants. Col exhibits high levels of heritable (CG and CHG) methylation, while *ddm1* shows similar levels of all methylation contexts, indicative of *de novo* methylation. This graph shows that there is a significant decrease in all levels of *Athila6* transcriptional start site methylation, but the most striking decrease in methylation is seen when AGO1 and AGO6 are not functioning. The *ddm1/ago9* WA background also seems to exhibit particularly low levels of methylation; however, the levels of methylation in *ddm1/ago9* SA are higher than in *ddm1/ago9* WA, and the total loss of function allele is a better representation of the true function of the protein.

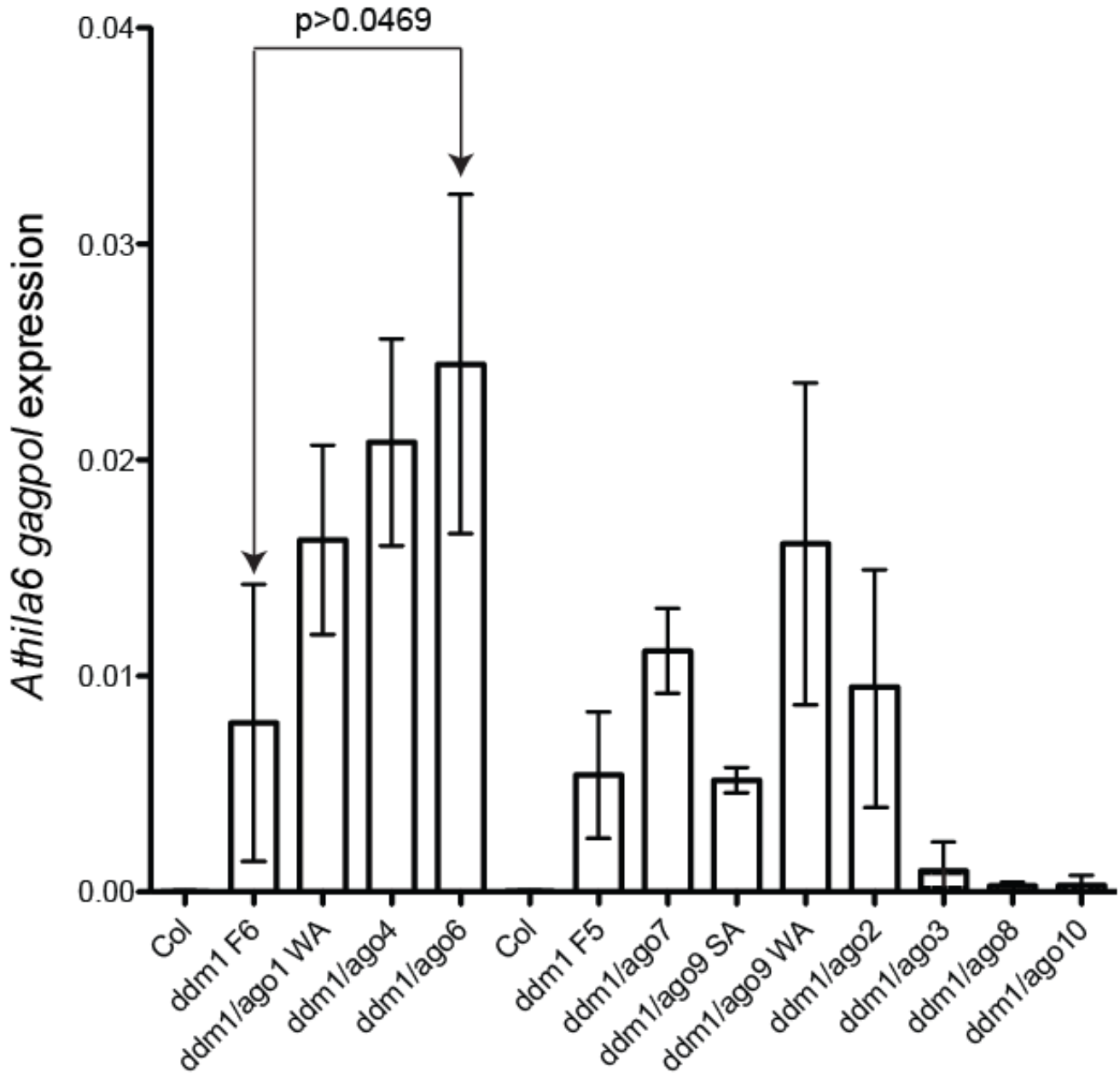


Figure 16: Quantitative PCR data of various *ddm1/ago* double mutants. The *gag/pol* region of the *Athila* transcripts is targeted because it is the region immediately downstream of the 5'LTR transcriptional start site. This way, the affect of start site methylation levels on *Athila* transcript levels can be observed. The only double mutant that shows a significant increase in *Athila* expression when compared to *ddm1* is *ddm1/ago6* with a p-value of .0469.

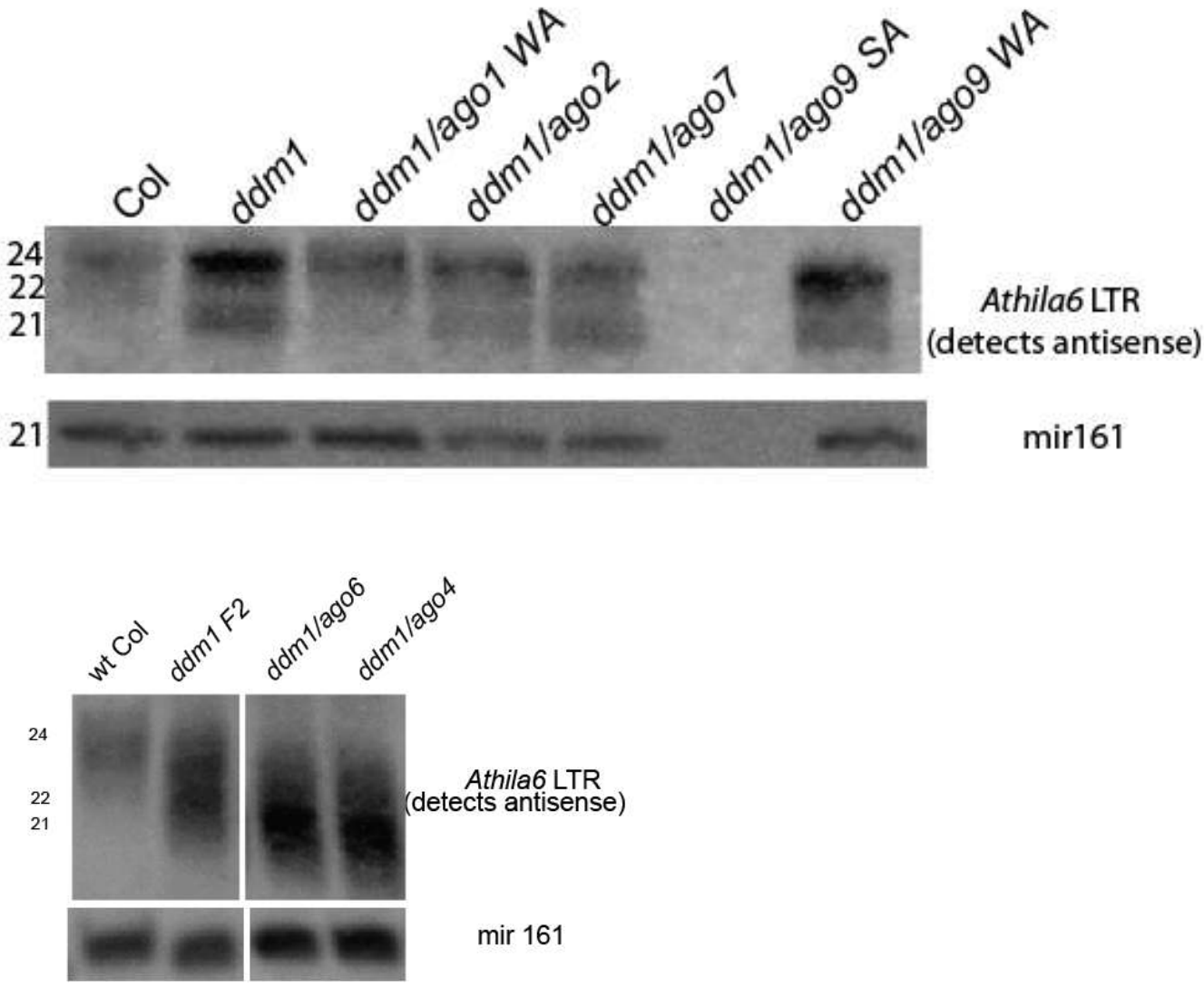


Figure 17: Small RNA Northern blots depicting levels of 21/22nt siRNAs and 24nt siRNAs in various *ddm1/ago* double mutant backgrounds. These blots show that 21/22nt and 24nt siRNAs levels are similar to those levels found in *ddm1*, except in *ddm1/ago1* where the production of 21/22nt siRNAs is lost. This data suggests that *ddm1/ago1* is involved in the biogenesis of 21/22nt siRNAs. This lack of 21/22nt siRNA production explains the decrease in methylation levels observed in the *ddm1/ago1* background. However, the decrease in methylation levels observed in the *ddm1/ago6* background cannot be explained by a decrease in siRNAs, suggesting that it is working to incorporate siRNAs for TE targeting rather than siRNA biogenesis. In addition, there appears to be a slight decrease in the levels of 21/22nt siRNAs in the *ddm1/ago2* background, suggesting that it could also play a role in the production of these siRNAs. The *ddm1/ago9* SA background failed to produce results (see the lack of a microRNA control band).

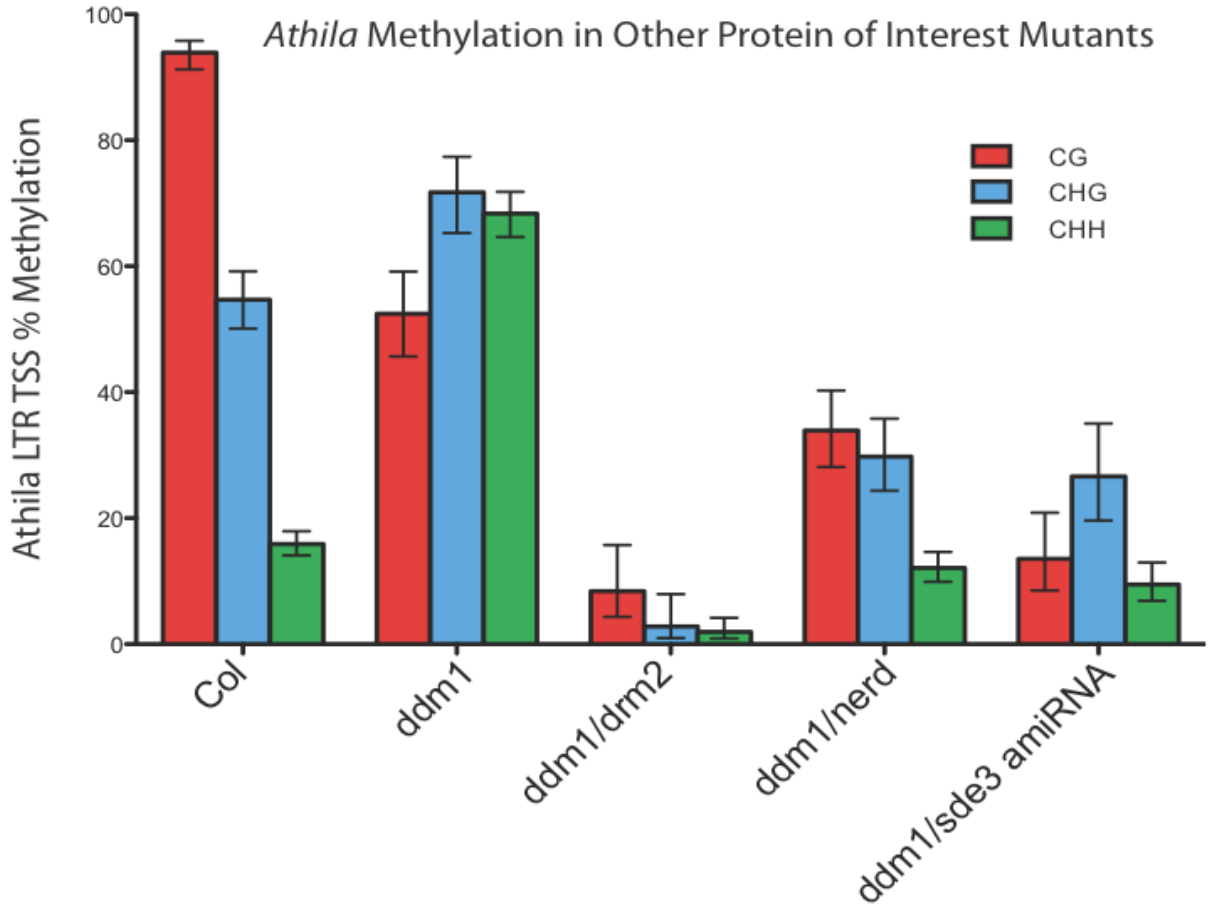


Figure 18: Bisulfite data from the other double mutants of interest. Levels of methylation decrease in all of the double mutants when compared to *ddm1*, suggesting that DRM2, NERD, and SDE3 could all potentially play a role in RDR6-RdDM. The loss of methylation in *ddm1/drm2* is the most dramatic, suggesting that it plays a particularly important role in the *de novo* methylation of TEs.

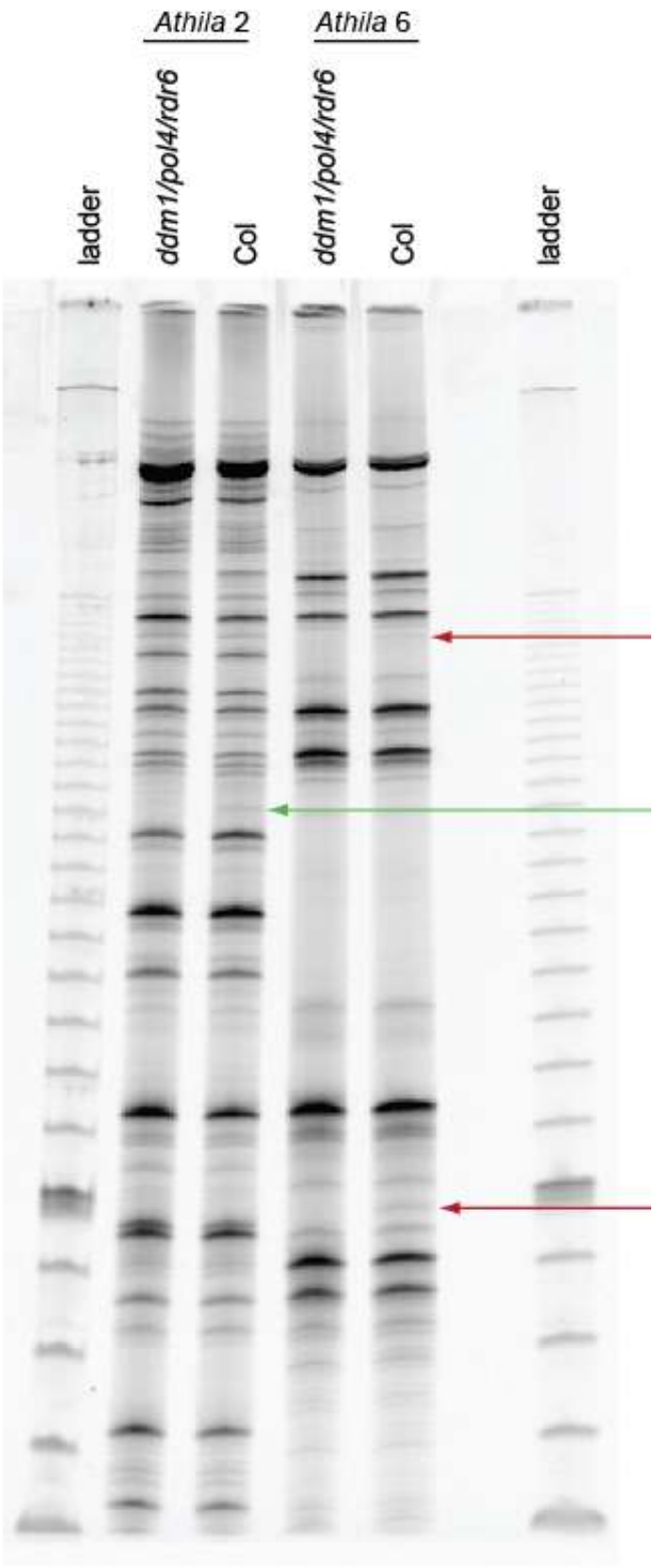


Figure 19: Transposable Element Display of *ddm1/pol4/rdr6* triple mutant compared to Col. Both *Athila6* and *Athila2* were analyzed in this assay. No new *Athila* transpositions were observed in the triple mutant when compared to the WT background. The arrows point to bands that exist in Col but not the triple mutant. These are most likely TEs that were recombined out of the genome during the generation of the triple mutant.

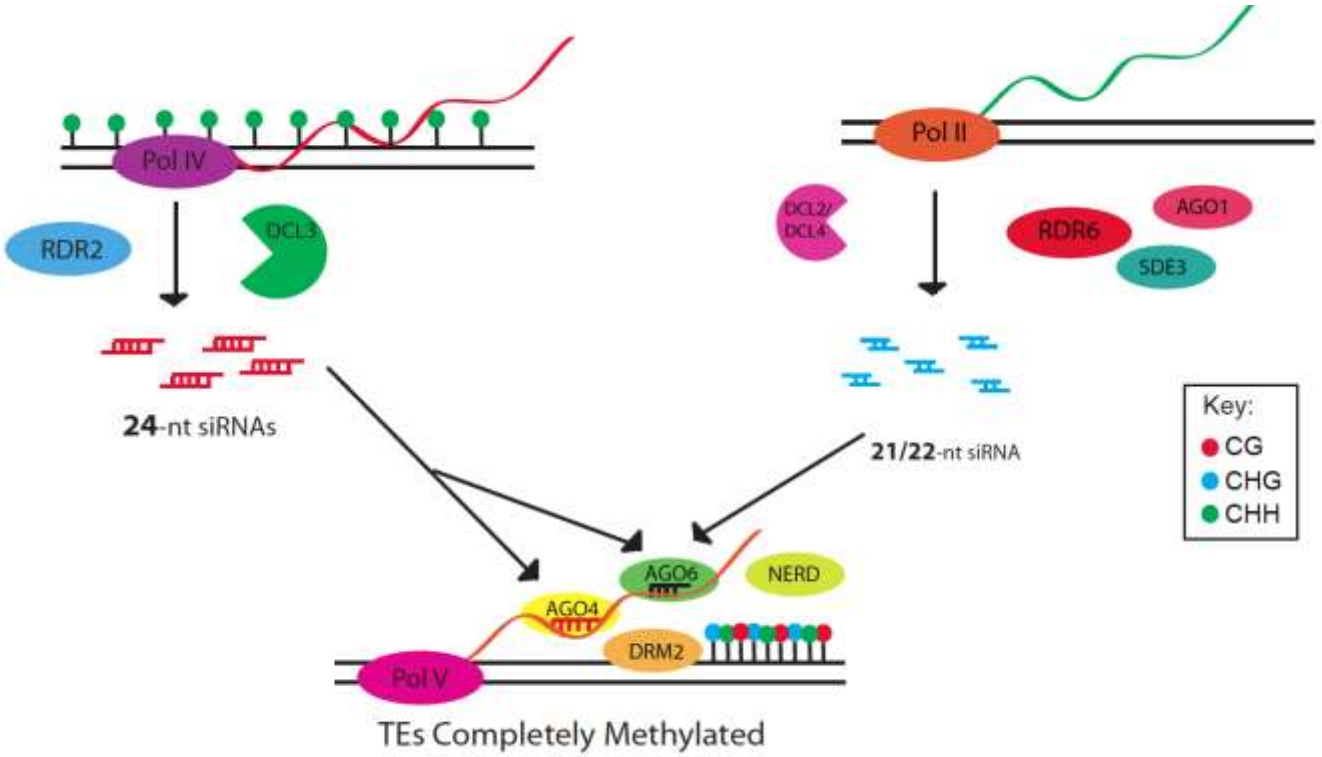


Figure 20: The Slotkin lab's current model of *de novo* methylation of transcriptionally active TEs. The previously known Pol4-RdDM generates 24nt siRNAs, which are incorporated into AGO4 or AGO6 and used to target TEs for *de novo* methylation. The recently discovered RDR6-RdDM pathway generates 21/22nt siRNAs, which are potentially utilizing AGO6 to target TEs for *de novo* methylation. Both of these pathways must be functional for TEs to obtain the same levels of *de novo* methylation observed in *ddm1*. At the intersection of these two pathways, it is suspected that DRM2 is the methyltransferase responsible for transferring the methyl groups onto the unmethylated TEs.

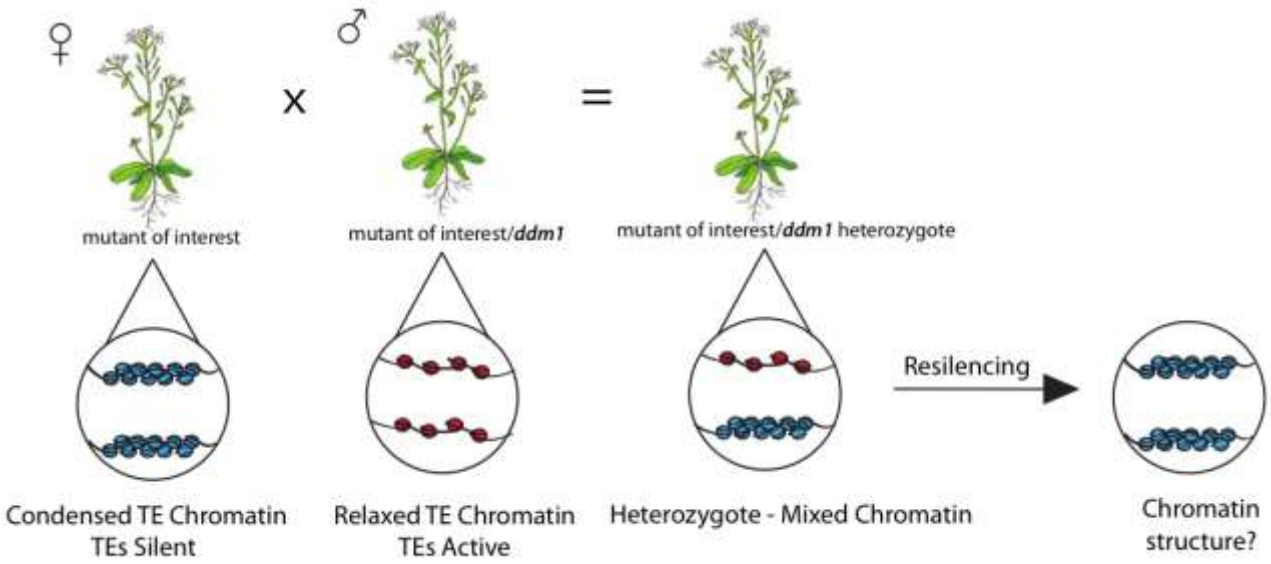


Figure 21: Depiction of the cross that was utilized to generate *ddm1* heterozygous lines. One parent which was homozygous for the particular mutant of interest and wild-type for DDM1 (silenced TEs) is crossed to another parent which was homozygous for the same mutant of interest and also homozygous for *ddm1* (active TEs). It has been previously shown that a wild-type DDM1 plant crossed to a *ddm1* mutant results in offspring that have fully resilenced *Athila6*, due to the activity of siRNAs which work though RdDM to reestablish the silencingj. The particular mutants of interest analyzed in these crosses are components of both RDR6-RdDM and Pol4-RdDM.

Arabidopsis Image Reference: http://www.physics.hmc.edu/research/ocm/plant.html

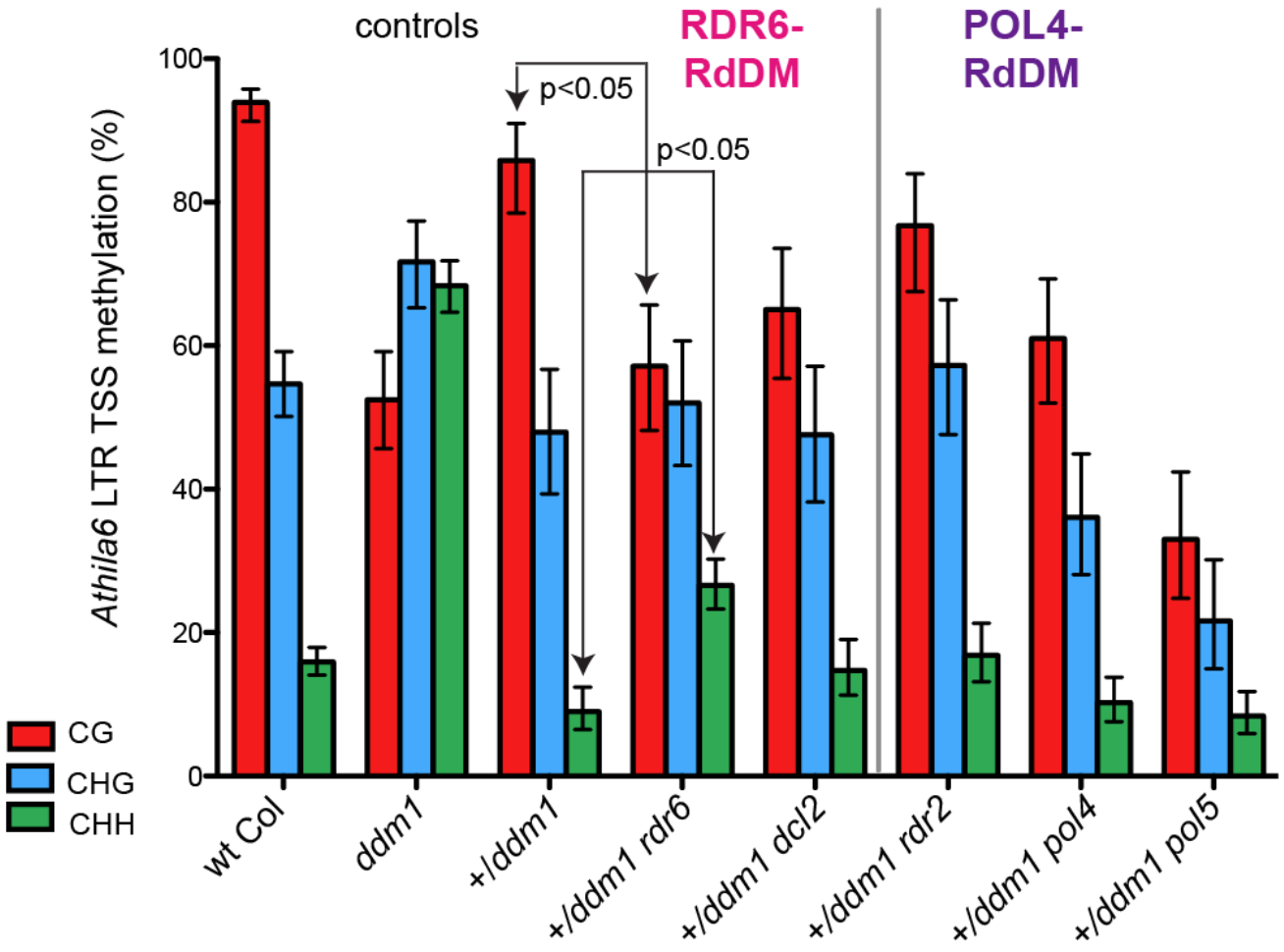


Figure 22: Results of the bisulfite sequencing for plants that are *ddm1* het; homozygous mutant for a particular protein of interest. The loss of components of RDR6-RdDM such as RDR6 and DCL2 results in the decrease of maintenance methylation (typically represented by CG and CHG methylation) when compared to *ddm1* hets, and the increase in *de novo* methylation (typically represented by an increase in CHH methylation). This is similar to what is seen when components of Pol4-RdDM are lost, supporting the idea that both of these pathways function to resilencing active TEs. Pol5 is suspected to work at the intersection between RDR6-RdDM and Pol4-RdDM because when it is lost, general methylation levels dramatically decrease.

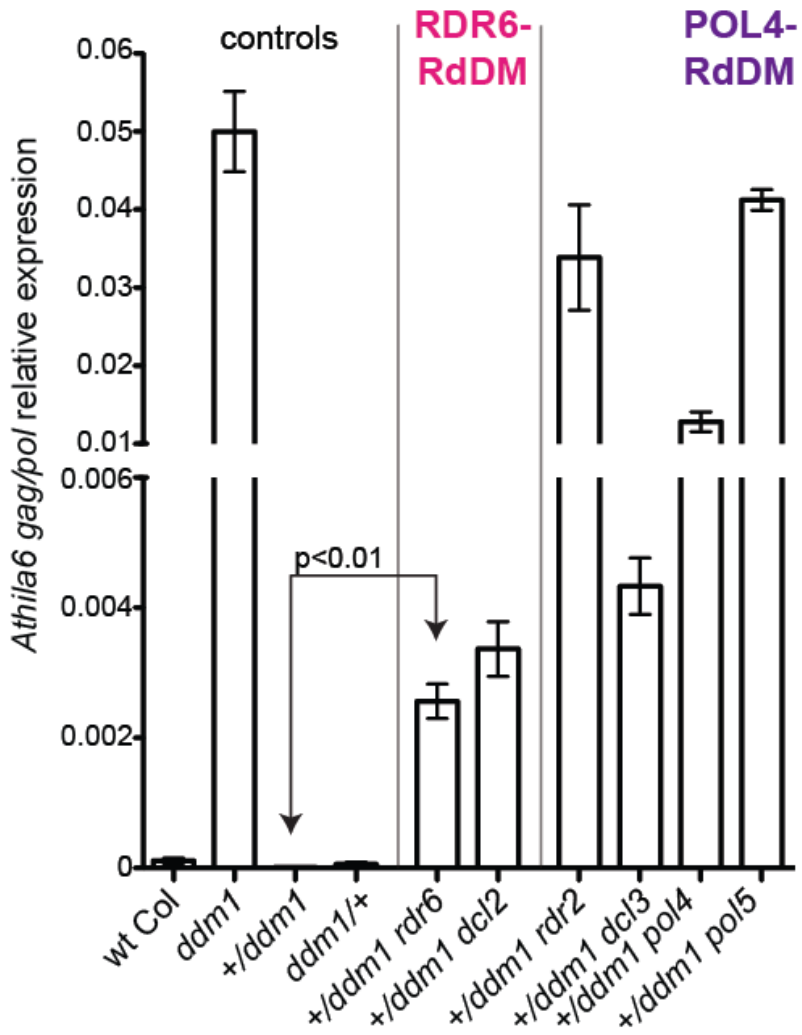


Figure 23: Quantitative PCR for plants that are *ddm1* het; homozygous mutant for a particular protein of interest. This data supports the previous bisulfite data in showing that TEs are still active and not properly resilenced when components of either RDR6-RdDM or Pol4-RdDM are lost.

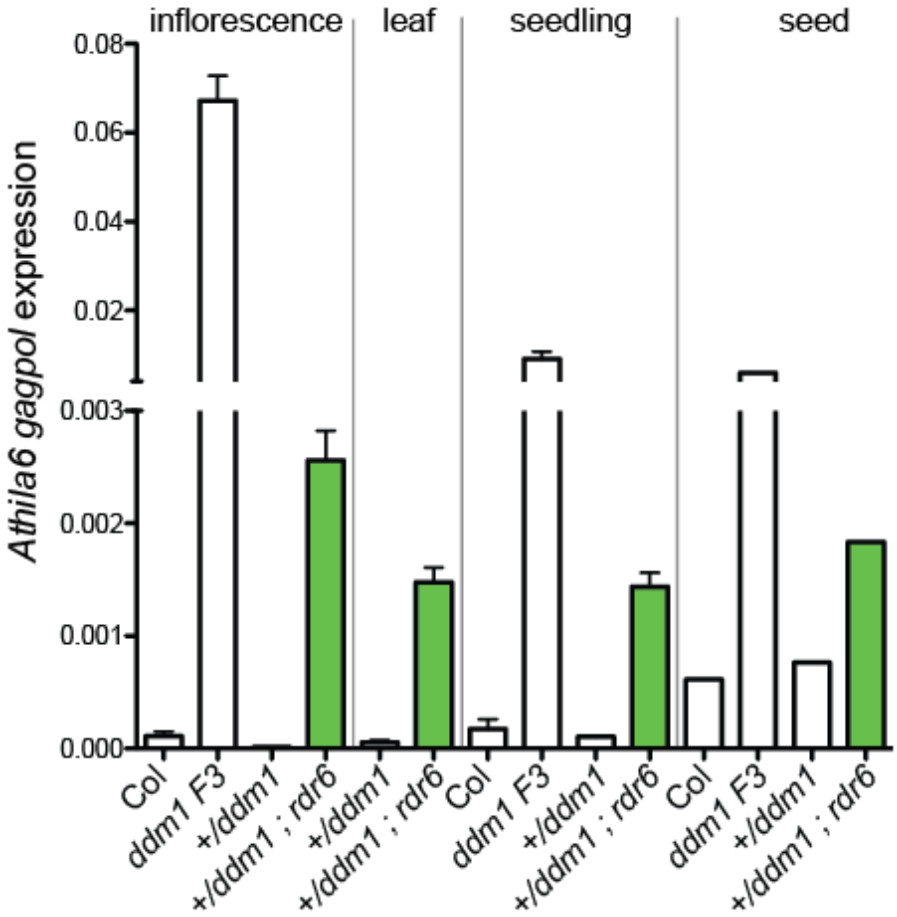


Figure 26: Quantitative PCR of seed tissue. Bioreps could not be performed in this assay due to a shortage of tissue. The expression of *Athila* in the *ddm1* het and *ddm1* het/ homo mutants of interest is not significantly similar. It can be concluded that the resilencing of TEs is not occurring in this tissue type.

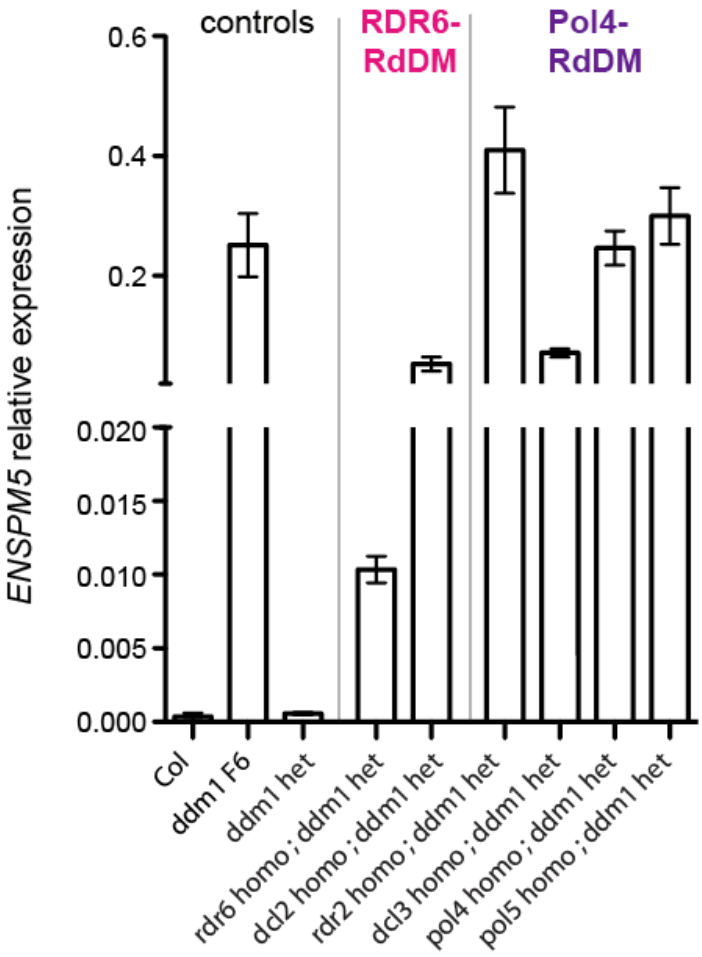


Figure 27: Quantitative PCR of *ENSPM5*. The expression levels in each one of the RDR6-RdDM and Pol4-RdDM backgrounds are similar to those observed in *Athila*. This supports the idea that RDR6-RdDM is required to fully resilence other TEs besides *Athila*.

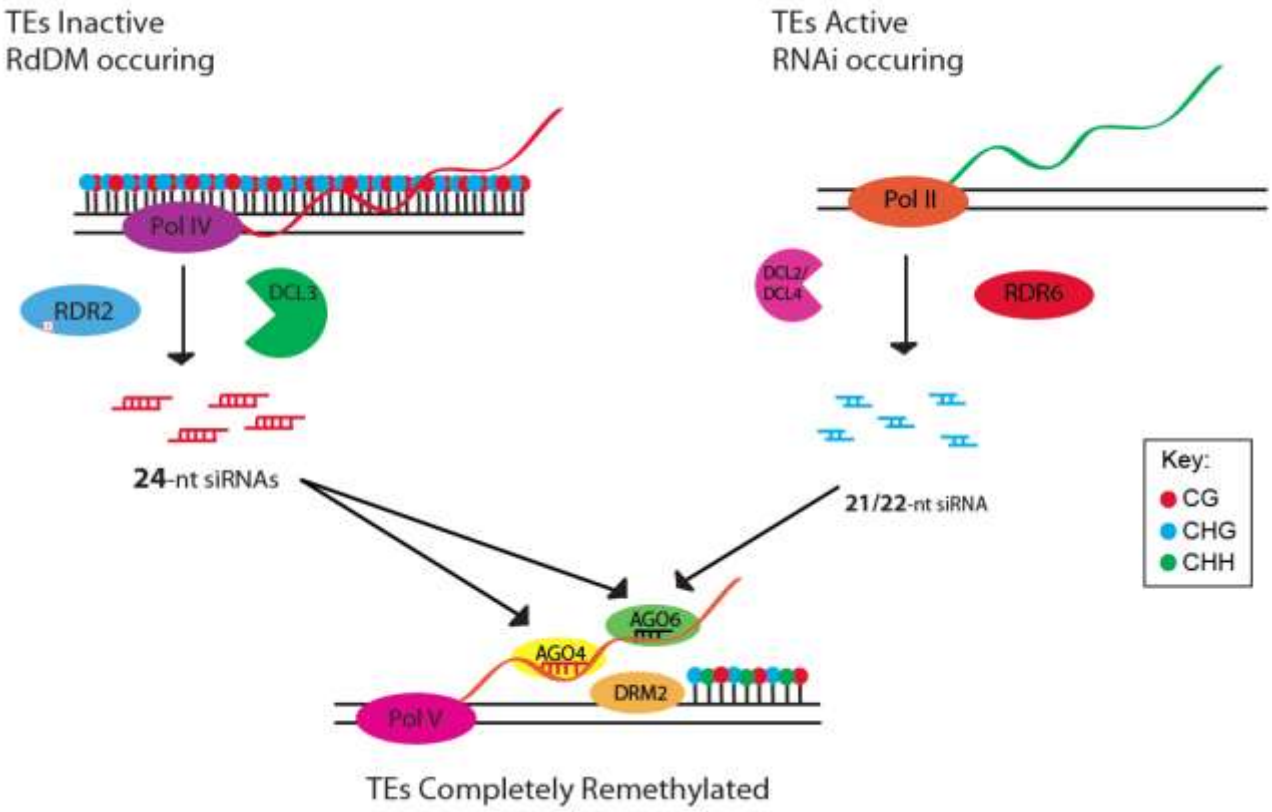


Figure 28: The Slotkin lab's current model of the RdDM-dependent resilencing of TEs. Pol4-RdDM is producing 24nt siRNAs from the epigenetically silenced TEs inherited from the DDM1 wild-type parent. RDR6-RdDM is producing 21/22nt siRNAs from the active TEs inherited from the *ddm1* parent. Together, both size classes of siRNAs are recruited to the unmethylated TEs to aid in the complete resilencing of these TEs.